DOCUMENT RESUME

ED 379 353 TM 022 748

AUTHOR	Shepard, Lorrie; And Others
TITLE	Second Report on Case Study of the Effects of Alternative Assessment in Instruction. Student Learning and Accountability Practices. Project 3.1. Studies in Improving Classroom and Local Assessments.
INSTITUTION	National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Sep 94
CONTRACT	R117G10027
NOTE	121p.; Papers presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Collected Works - General (020) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Accountability; Case Studies; *Educational Assessment; *Elementary School Students; Grade 3; Instructional Effectiveness; *Learning; Primary Education; Program Evaluation; *Student Attitudes; Student Evaluation; *Test Construction; Test Use
IDENTIFIERS	*Performance Based Evaluation

ABSTRACT

Three papers are presented that summarize current project findings from a study of the actual effects of introducing new forms of assessment at the classroom level. All focus on aspects of performance assessment as an alternative to traditional assessments. "Effects of Introducing Classroom Performance Assessments on Student Learning" by Lorrie A. Shepard, and others, examines effects of performance assessment on the learning of third graders in 13 classrooms. "'How Does my Teacher Know What I Know?' Third Graders' Perceptions of Math, Reading, and Assessment" by Kathryn H. Davinroy, Carribeth L. Bliem, and Vicky Mayfield uses interviews with students in the classrooms of the larger study to explore student ideas and attitudes. "How 'Messing About' with Performance Assessment in Mathematics Affects What Happens in Classrooms" by Roberta J. Flexner reviews work with the teachers of the study's classes. Eighteen tables and six figures in the three papers present study findings. (Contain 77 references in all.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 379 353

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – September 1994

Project 3.1 Studies in Improving
Classroom and Local Assessments

Second Report on Case Study of the Effects of Alternative Assessment in Instruction, Student Learning and Accountability Practices

Lorrie Shepard, Project Director
CRESST/University of Colorado at Boulder

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM022748

BEST COPY AVAILABLE

The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

CONTENTS

PREFACE v

EFFECTS OF INTRODUCING CLASSROOM PERFORMANCE ASSESSMENTS ON STUDENT LEARNING 1

Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert, Scott F. Marion, Vicky Mayfield, and Timothy J. Weston

"HOW DOES MY TEACHER KNOW WHAT I KNOW?" THIRD GRADERS' PERCEPTIONS OF MATH, READING, AND ASSESSMENT 31

Kathryn H. Davinroy, Carribeth L. Bliem, and Vicky Mayfield

HOW "MESSING ABOUT" WITH PERFORMANCE ASSESSMENT IN MATHEMATICS AFFECTS WHAT HAPPENS IN CLASSROOMS 67

Roberta J. Flexer

PREFACE

The current intense interest in alternative forms of assessment is based on a number of assumptions that are as yet untested. In particular, the claim that authentic assessments will improve instruction and student learning is supported only by negative evidence from research on the effects of traditional multiple-choice tests. Because it has been shown that student learning is reduced by teaching to tests of low-level skills, it is theorized that teaching to more curricularly defensible tests will improve student learning (Frederiksen & Collins, 1989; Resnick & Resnick, 1992). In our current research for the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) we are examining the actual effects of introducing new forms of assessment at the classroom level.

Derived from theoretical arguments about the anticipated effects of authentic assessments and from the framework of past empirical studies that examined the effects of standardized tests (Shepard, 1991), our study examines a number of interrelated research questions:

1. What logistical constraints must be respected in developing alternative assessments for classroom purposes? What are the features of assessments that can feasibly be integrated with instruction?
2. What changes occur in teachers' knowledge and beliefs about assessment as a result of the project? What changes occur in classroom assessment practices? Are these changes different in writing, reading, and mathematics, or by type of school?
3. What changes occur in teachers' knowledge and beliefs about instruction as a result of the project? What changes occur in instructional practices? Are these changes different in writing, reading, and mathematics, or by type of school?
4. What is the effect of new assessments on student learning? What picture of student learning is suggested by improvements as measured by the new assessments? Are gains in student achievement corroborated by external measures?
5. What is the impact of new assessments on parents' understandings of the curriculum and their children's progress? Are new forms of assessment credible to parents and other "accountability audiences" such as school boards and accountability committees?

The enclosed three papers, which were presented at the 1994 annual meeting of the American Educational Research Association, summarize current project findings.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *73*, 232-238.

EFFECTS OF INTRODUCING CLASSROOM PERFORMANCE ASSESSMENTS ON STUDENT LEARNING[1,2]

**Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert,
Scott F. Marion, Vicky Mayfield, and Timothy J. Weston**

CRESST/University of Colorado at Boulder

Arguments favoring the use of performance assessments make two related but distinct claims. Performance assessments are expected first to provide better measurement and, second, to improve teaching and learning. Although any measuring device is corruptible, performance measures have the potential for increased validity because the performance tasks are themselves demonstrations of important learning goals rather than indirect indicators of achievement (Resnick & Resnick, 1992). According to Frederiksen and Collins (1989), Wiggins (1989), and others, performance assessments should enhance the validity of measurement by representing the full range of desired learning outcomes; by preserving the complexity (and ambiguity) of disciplinary knowledge domains and skills; by representing the contexts in which knowledge must ultimately be applied; and by adapting the modes of assessment to enable students to show what they know. The more assessments embody authentic criterion performances, the less we have to worry about drawing inferences from test results to remote constructs.

The expected positive effects of performance assessments on teaching and learning follow from their substantive validity. If assessments capture learning expectations fully, then when teachers provide coaching and practice to improve scores, they will directly improve student learning without corrupting the meaning of the indicator. Resnick and Resnick (1992),

---

[1] Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1994

[2] We thank the Maryland Department of Education for allowing us to use tasks from the Maryland School Performance Assessment Program as outcome measures for the study. We also thank the Riverside Publishing Company for permission to use portions of the 2nd-grade ITBS as a premeasure.

Frederiksen and Collins (1989), and Wiggins (1989) all argue that it is natural for teachers to work hard to prepare their students to do well on examinations that matter. Rather than forbid "teaching to the test," which is impossible it is preferable to create measures that will result in good instruction even when teachers do what is natural. The reshaping of instruction toward desirable processes and outcomes is expected to occur both indirectly, as teachers individually imitate assessment tasks in a variety of ways, and directly, because criteria for judging performances will be shared explicitly.

These anticipated benefits of performance assessments are conceptually grounded and supported by evidence documenting the negative effects of traditional, standardized testing. Under conditions of high-stakes accountability pressure, it has been demonstrated that teachers align instruction with the content of basic skills tests, often ignoring science and social studies and even untested objectives in reading and mathematics. Furthermore, instruction on tested skills comes to resemble closely the format of multiple-choice tests, with students learning to recognize right answers rather than generating their own problem solutions (Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard, 1991; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990). Such measurement-driven instruction has been harmful to learning as evidenced by the decline in higher order thinking skills on the National Assessment of Educational Progress during the 1980s and by the failure of accountability test score results to generalize when students are retested using less familiar formats (Flexer, 1991; Hiebert, 1991; Koretz, Linn, Dunbar, & Shepard, 1991).

Thus it is the obverse case that has been "proven." Teaching to standardized tests harms both teaching and learning. Advocates of performance assessments assume, therefore, that parallel mechanisms will work to produce positive effects once limited tests are replaced by more desirable measures. However, to date little research has been done to evaluate the actual effects of performance assessments on instructional practices or on student learning. Although some extreme views hold that authentic performance measures are valid by definition and will automatically produce salutary effects, we would argue in contrast that the effects of performance assessments should be evaluated empirically following a program of inquiry closely parallel to the studies undertaken to examine the effects of standardized tests. We concur with Linn, Baker, and Dunbar (1991) that validity criteria for alternative assessments should address intended and unintended effects as well as more substantive features such as cognitive complexity, content quality and comprehensiveness, generalizability of knowledge from assessed to unassessed tasks, and the like. Although we are committed to performance assessments on conceptual grounds, their demonstrated effects on teaching and learning remain an open question.

The purpose of the present study was to examine the effects of performance assessments on student learning. If standardized tests are removed and teachers begin to use performance assessments as part of regular instruction, will student performance on independent measures of achievement be improved? Note that some arguments favoring the use of performance assessments to leverage educational reform presume that the high-stakes accountability pressures would still be needed to drive instructional change. Other advocates focus more on the informational and feedback effects of classroom-embedded assessments. In this study, we adopted the second perspective. We were interested in the effects of using assessments as part of instruction but without the incentives and context created by an externally mandated system.

A year-long project was undertaken to help teachers in 13 third-grade classrooms begin to use performance assessments as a part of regular instruction in reading and mathematics. Other parts of the research project focused on changes in teachers' beliefs and practices about summaries and expository text in reading (Borko, Davinroy, Flory, Hiebert, 1994; Davinroy & Hiebert, 1993); on changes in teachers' beliefs about assessment and instruction in mathematics (Flexer, Borko, Cumbo, Marion, & Mayfield, 1994); on parent attitudes toward performance assessments (Shepard & Bliem, 1993); and on student understandings of how teachers "know what they know" (Davinroy, Bliem, & Mayfield, 1994). Here research questions are focused on student achievement in reading and mathematics. Did students learn more or develop qualitatively different understandings because performance assessments were introduced into classrooms? Achievement results were compared both to the performance of third-grade students in the same schools the year before and to third-grade performance in matched control schools.

Study Methods

Setting

The study was conducted in a working-class and lower-to-middle-class school district on the outskirts of Denver, Colorado. The district was selected in part because of the willingness of central office administrators to participate and in part because of its ethnically diverse student population. In the 1980s the district was known for its extensive mastery learning and criterion-referenced testing system, but more recently curriculum guidelines in language arts and mathematics were revised to reflect more constructivist conceptions of these disciplines, consistent with national standards (Anderson, Hiebert, Scott, & Wilkinson, 1985; National Council of Teachers of Mathematics, 1989).

We wanted teachers to be free to implement performance assessments and to make concomitant changes in instruction without worrying about how their students would do on the standardized test normally administered every April. Therefore, a requirement of participation was that the district be willing to apply to the state for a 2-year waiver from standardized testing in the three schools selected to participate. As part of its procedures to grant the waiver the state required, in turn, that approvals be obtained from the school board, the district accountability committee, the teachers' union, and parent accountability committees in each of the participating schools.

Sample and Research Design

Third grade was selected as the target grade level because district CTBS testing occurs at Grades 3 and 6 (but not all sixth grades are in elementary schools). Because of the amount of time and effort that would be required of teachers, volunteer schools were sought. Third-grade teachers had to make a commitment as a team with the support of their principal and parent accountability committee. Ten schools sent representatives to a workshop where the study's purpose and methods were explained. We accepted the three schools that completed the formal application to participate. In the 1992-93 study year there were 13 third-grade classrooms in the three schools combined involving approximately 335 third graders.

Three control schools were identified to be used for comparison when analyzing teachers' beliefs and parents' opinions as well as students' achievement. The control and participating schools were matched on free and reduced lunch percentages, percentage of minority children, and other knowledge of neighborhood similarities such as type of housing. Data in Table 1 show the socioeconomic differences among the three participating schools as well as their matches to control schools. Note that the implementation year of the project was 1992-93. Therefore, site selection and baseline testing occurred in the spring of 1992; data available for school matching had been gathered spring 1991.

CTBS achievement test data for participating and control schools are shown separately in Table 2. As part of the matching process, we found that it was impossible to match schools on both socioeconomic factors and 1991 CTBS scores because they diverged too much. This was unusual. In our experience in other studies, test scores and socioeconomic indicators usually correspond closely enough that it is possible to select schools that are the same on both. Because we could not know whether sharp differences in achievement scores meant more able populations, more able teaching, or more test-score inflation in the candidate control schools, we elected to match only on socioeconomic data. However, subsequent to the selection process, we administered our own baseline achievement measures in reading and mathematics which confirmed the superior performance of third graders in control schools in the year before the study began.

The research design called for two separate comparisons. Outcome measures in reading and mathematics selected for administration in May 1993 were also administered as baseline measures in May 1992. In addition, premeasures appropriate to entering third graders were administered in September 1992 and used as covariates to evaluate 1993 outcomes.

Assessment Project "Intervention"

The intention of the project was not to introduce an already-developed curriculum and assessment package. Rather, we proposed to work with teachers to help them develop (or select) performance assessments congruent with their own instructional goals. Faculty researchers included Roberta Flexer, an expert in mathematics, Elfrieda Hiebert, an expert in reading, Hilda Borko, whose specialty is teacher change, and Lorrie Shepard, an assessment expert. We met with teachers for planning meetings in spring

Table 1

Socioeconomic Characteristics of Participating and Control Schools

	Participating schools			Control schools		
	1	2	3	1	2	3
Free and reduced lunch	61%	9%	6%	55%	13%	3%
Percent minority	37%	16%	14%	45%	19%	10%
Student turnover	27%	7%	11%	30%	11%	10%

Table 2

Grade 3 Mean CTBS Scores in Reading and Mathematics for Participating and Control Schools

	Participating schools			Control schools		
	1	2	3	1	2	3
5-Year Average (1987-91)						
Total reading	47.8	48.8	52.7	48.9	50.4	54.7
Total mathematics	52.5	47.5	51.3	49.3	60.9	58.1
1991						
Total reading	47.0	43.0	47.0	47.0	54.0	56.0
Total mathematics	54.0	52.0	50.0	50.0	67.0	63.0
1992						
Total reading	44.6	51.9	55.5	43.1	54.4	57.2
Total mathematics	53.9	53.8	62.8	47.5	66.5	68.1
1993						
Total reading	N/A[a]	N/A[a]	N/A[a]	49.6	47.0	57.2
Total mathematics	N/A[a]	N/A[a]	N/A[a]	57.1	57.1	65.3

[a] Participating schools were exempt from CTBS testing in 1993 as a condition of the study.

1992 and September 1992. Then we met for weekly afterschool workshops for the entire 1992-93 school year, alternating between reading and mathematics so that subject matter specialists could rotate among schools.

Because the district had newly developed curriculum frameworks consistent with emerging national standards in reading and mathematics, and because teachers had volunteered to participate in the project, we assumed that their views about instruction would be similar to those reflected in the district curriculum and therefore similar to our own. What we learned later was that not all teachers were true volunteers; some had been "volunteered" by their principals or had acquiesced to pressure from the rest of the third-grade team. More importantly, for understanding the substantive character of the project, even some teachers who were willing and energetic project participants were happy with the use of basal readers and chapter tests in the math text and were not necessarily familiar with curricular shifts implied by the new district framework in mathematics.

Although dissonance between researchers' and teachers' views about subject matter instruction was sometimes acknowledged and joked about in workshops, for the most part researchers avoided confrontations about differences in beliefs and did not propose radical changes in instruction. Faculty experts worked to suggest possible reading and mathematics activities that addressed teachers' goals but that departed from a strictly skills-based approach. For example, we refused to consider having timed tests on math facts as part of project portfolios, but in other ways we conformed to teacher-identified goals.

At the start of the year, teachers said that they wanted to "start small" but then selected as goals meaning-making and fluency in reading, and understanding of place value, addition and subtraction, and multiplication, as the foci for the project. In the fall, for reading, teachers learned to use running-records to assess fluency for below-grade-level readers. Written summaries were used to assess comprehension; but for some teachers summaries became an end in themselves (Borko, et al., 1994). Project activities included the development of rubrics to score written summaries. In the spring, ideas about meaning-making and written summaries were extended to expository texts.

In mathematics, teachers made extensive requests, throughout the year, for materials and ideas for teaching the topics of the third-grade curriculum, for example, place value, addition, geometry, and probability. Materials that addressed these topics from a problem-oriented and hands-on approach were distributed to all three schools to use in both instruction and assessment. Teachers were offered nonroutine problems from which to select a number to try with their classes. Some problems required students to explain their solutions; others required students to analyze and explain an incorrect step or computation in a buggy problem. Materials were also distributed for making and using base 10 blocks for modeling numbers and operations. Some teachers had not previously worked with place-value mats or manipulatives and introduced them for the first time. Discussions at weekly meetings included dialogue about using materials for both instruction and assessment, making observations and how to keep track of them, analyzing student work, and developing rubrics for scoring problem solving and explanations.

Outcome Measures and Covariates

For obvious reasons, we did not wish to use a multiple-choice standardized test to measure the project's effects. At the same time, a compendium of performance tasks used throughout the project would also not be a fair outcome measure. We are grateful to the Maryland State Department of Education for allowing us to use portions of their 1991 performance assessments in reading and mathematics as outcome measures for the study. Although the Maryland assessments are still relatively test-like compared to week-long projects that students might do, they are markedly different from traditional tests. The tasks provide sufficient structure and support so that students in the baseline year and in control schools could understand what they were being asked to do, but they are sufficiently open-ended that students had to produce answers to show what they knew. In literacy, students read extended stories and informational texts in a separate reading book and then wrote responses about what they read, completed tables, drew story webs, and so forth. In mathematics, tasks involved a series of problems all related to the same information source or application. Students had to solve problems that involved identifying patterns, estimating as well as computing, using calculators, extending tables, and explaining how they got their answers. Because we were limited to only four 1-hour sessions to administer our outcome measures, we used only a sample of tasks from the Maryland assessments.

We wanted to be sure to assess a range of skills in mathematics. Therefore, we used three tasks from the Maryland assessment in one 1-hour session, but also used a portion of an alternative measure in mathematics developed for another study (Koretz et al., 1991). This test consisted of 15 short-answer and multiple-choice items that assess problem solving in, and conceptual understanding of, functions and relations, patterns, whole-number operations, probability, and data and graphs. Problem-types included application and nonroutine problems.

Covariate measures were needed for entering third graders to assess their initial abilities in reading and mathematics. In reading, portions of a Silver, Burdett and Ginn 2/3 Reading Process Test and 2/3 Skills Progress Test were used with permission from the publisher. After reading a 13-page story and pictures book, students responded to questions by checking answers (more than one answer could be correct) and also writing responses. Students also read two page-long passages and responded to comprehension questions by circling the correct answer. In mathematics, open-ended problems were developed to measure students' ability to discern patterns and number relations. This subtest was combined with three subtests from the second-grade level of the Iowa Tests of Basic Skills covering math concepts, estimation, and data interpretation. We are grateful to Riverside Publishing Company for allowing us to reproduce portions of their test for use in the study. The reading and math covariates were each administered in 1-hour sessions on separate days.

Scoring and Reliability

All of the measures used in the study required scoring of open-ended student responses. In particular, the Maryland assessment tasks required scorers to make subjective judgments about the quality of student answers. Therefore, these instruments received the greatest scrutiny in our reliability studies. Scorers worked from the scoring guides provided by the Maryland School Performance Assessment Program with slight modifications made by the respective subject matter experts. Day-long training sessions were held in summer 1992, and again in 1993, to ensure that scorers were familiar with the scoring rules and able to apply them to the full range of students' responses.

Interrater reliability was assessed both within year (are all of the scorers rating consistently?) and between years (were the scoring rules implemented consistently in 1992 and 1993?). For the within-year studies, three student booklets in reading and three in mathematics were chosen at random from each classroom. This resulted in more than a 10% sample with 55 to 60 out of 500 booklets being rescored. Booklets were scored independently by the scorer-trainer. Three other raters were then compared one at a time and then in aggregate to this standard rater. Pearson correlations between total scores assigned by other raters and by the standard rater were quite high in both years for both reading and mathematics; values ranged from .96 to .99. The Maryland reading measure was composed of 61 scored "items" or task subparts; the Maryland mathematics measure had 31 scorable entities. The high correlations between raters simply mean that with sufficient numbers of task subscores, raters can rank students quite accurately.

A truer picture of the effect of rater agreement on total scores is provided by the data in Tables 3 and 4. On individual items requiring a subjective judgment, raters might differ by only one point in how they scored the item. However, these discrepancies could accumulate across items. The data in Tables 3 and 4 show how often raters agreed completely with the standard rater on total score and how often they differed by four or more points in reading or two or more points in mathematics. Within years, raters agreed on total score within one or two points for 97% or 98% of cases in reading and for 90% to 91% of cases in mathematics. These agreement rates are respectable

Table 3

Percentage of Scorer Agreement on Maryland Reading Assessment Total Score

	Within year 1992	Within year 1993	Between year 1992/1993
Complete agreement	33.3%	28.6%	12.2%
Agreement within ± 2 points (± .17 *SD*)	97.8%	96.5%	45.6%
Agreement within ± 4 points (± .34 *SD*)	100.0%	100.0%	71.9%
Range of differences	-2% to +4%	-3% to +2%	-5% to +9%

Note. Agreement is based on the comparison between each rater's judgment of total student score and the independent rater's judgment of student scores. ± 4 points was used in the reading analyses because the total number of possible points was 61 compared to 31 in the mathematics assessment. In standard deviation (*SD*) units, differences of 2 and 4 points in reading are roughly comparable to 1 and 2 points in mathematics.

Table 4

Percentage of Scorer Agreement[a] on Maryland Mathematics Assessment Total Score

	Within year 1992	Within year 1993	Between year 1992/1993
Complete agreement	30.4%	29.0%	14.3%
Agreement within ± 1 point (± .15 *SD*)	75.0%	64.5%	54.0%
Agreement within ± 2 points (± .31 *SD*)	91.0%	90.3%	79.4%
Range of differences	-4% to +3%	-4% to +3%	-3% to +4%

[a] Agreement is based on the comparison between each rater's judgment of total student score and the independent rater's judgment of student scores.

for subjectively scored instruments but nonetheless introduce noise into the evaluation of effects.

To check for consistency of scoring across years, test booklets from 1992 were "seeded" into 1993 classroom sets without scorers being aware of which booklets were being rescored. A total of 57 booklets were rescored in both mathematics and reading. As seen in Tables 3 and 4, the between-year agreements were not so high as the within-year agreements. In mathematics, 79% of total scores were within 2 points of the score assigned to the same booklet the year before. In reading, 72% were within 4 points (which is comparable in standard deviation units to a 2-point difference on the mathematics assessment). The between-years analysis also revealed some systematic biases with raters tending to become more stringent in 1993 than raters had been in 1992. In reading there was an average mean score shift downward for the 57 1992 booklets rescored in 1993 of 2.47 points. In math the greater stringency created a downward shift of .25 points. Because the reading score shift was both statistically and practically significant, 1993 reading scores were adjusted to correct for the systematic bias. Average biases varied for individual raters from 1.13 to 3.63, all in the direction of greater stringency; these specific corrections were applied to the sets of booklets scored by each rater.

Internal consistency coefficients provide another indicator of the psychometric adequacy of research instruments. Coefficients calculated on the entire sample are shown in Table 5 for the covariates and for both the 1992

Table 5

Internal Consistency Coefficients (Cronbach's Alpha) for Outcome and Covariate Measures

	1992		1992	
	n	alpha	*n*	alpha
Maryland reading	458	.90	458	.90
Covariate reading		n/a	458	.74
Maryland mathematics	487	.84	523	.83
Alternative mathematics	487	.78	524	.80
Covariate mathematics		n/a	454	.85

and 1993 administrations of the outcome measures. Although low coefficients could mean either poor reliability or task-item heterogeneity, high values provide assurance that summary scores are reliable and reasonably consistent measures of student performance.

Results

Outcome measures for 1993 were analyzed in two ways, first in comparison to 1992 baseline administrations of the same measures, and then in relation to control group outcomes using analysis of covariance. To make it easier to follow the logic of the two analyses, results are reported separately in Tables 6 and 7. The 1993 data are repeated in both tables, although subjects without pretest data were deleted from the analysis of covariance (Table 7). Then the analyses are presented graphically in Figures 1 and 2 for reading and mathematics respectively.

Overall, the predominant finding is one of "no-difference" or no gains in student learning following from the year-long effort to introduce performance assessments. Although we argue subsequently that we see qualitative changes in student performance and that the small gain in mathematics is real, honest discussion of project effects must acknowledge that any benefits are small and ephemeral. For example, improvements show up in some project-teachers' classrooms but not in all, and the largest gain from 1992 to 1993 for the participating schools combined, which occurred on the Maryland mathematics assessment, produced an effect size of only .13.

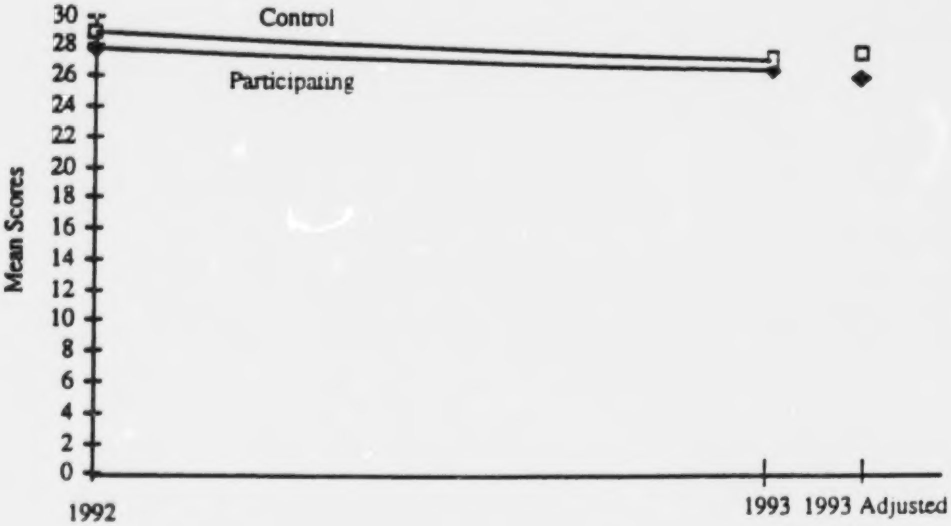


Figure 1
Maryland Reading Assessment Mean Scores for Participating and Control Schools.

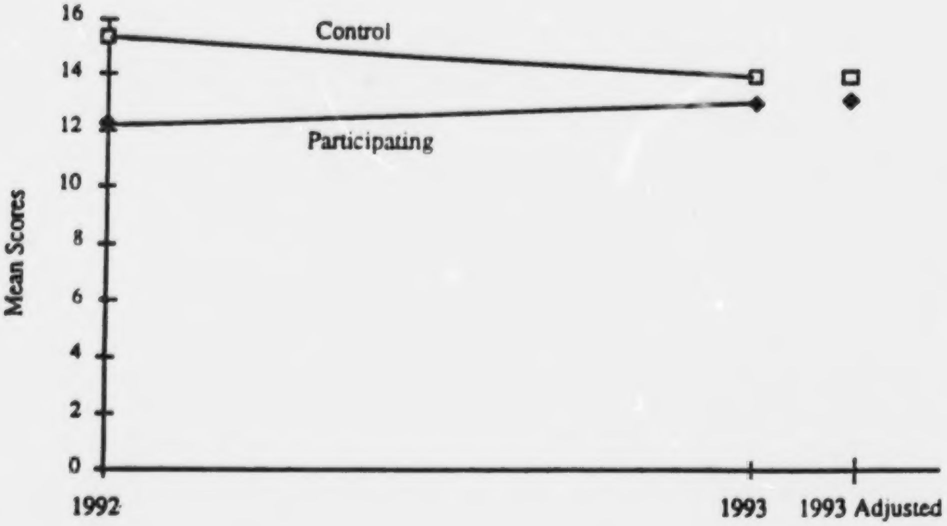


Figure 2
Maryland Mathematics Assessment Mean Scores for Participating and Control Schools

Table 6

1992 vs. 1993 Comparisons in Reading and Mathematics for Participating and Control Schools

	1992 Mean (*n*)	1993 Mean (*n*)	92-93 Mean difference	1992 Pooled w/in school *SD*	ES[a] of difference
Maryland reading total					
Participating	27.7 (290)	26.1 (305)	-1.6	11.7	-.14
Control	28.9 (210)	26.5 (228)	-2.4		-.21
Maryland math total					
Participating	12.2 (288)	13.0 (305)	0.8	5.94	.13
Control	15.3 (210)	13.6 (231)	-1.7		-.29
Alternative math total					
Participating	12.7 (288)	12.9 (305)	0.2	3.5	.06
Control	13.3 (208)	13.5 (229)	0.2		.06

[a] Effect size calculations are based on pooled within-school 1992 standard deviations using both participating and control group schools.

In reading there were no significant differences between 1992 and 1993 results or between participating and control schools. Both groups of schools appeared to lose ground slightly (.9 and 1.9 points respectively).

In mathematics the alternative test also showed no effects. However, the Maryland assessment in mathematics, which requires students to do more extended problems and explain their answers, did show a positive gain in the participating schools. We interpret this change, albeit small, as a "real" gain based on the following arguments. First, CTBS results for 1993 showed declines districtwide and in two of the control schools. Against a backdrop of declining achievement, slight gains in the participating schools are more impressive. Although the populations of the participating and control schools are quite similar as evidenced by socioeconomic variables and pretest

Table 7

1993 Outcome Comparisons Between Participating and Control Schools With and Without Covariance Adjustments

	1993 Mean	Sept. 1992 pretest[a]	May 1993 adjusted mean[b]
Maryland reading total			
Participating	26.8	11.7	26.2
Control	27.0	10.8	27.9
Difference	-0.2	0.9	-1.7
Maryland math total			
Participating	13.0	19.8	13.1
Control	13.9	20.4	13.8
Difference	-0.9	-0.6	-0.7
Alternative math total			
Participating	13.0	19.8	13.0
Control	13.8	20.4	13.7
Difference	-0.8	-0.6	-0.7

[a] There was one mathematics pretest and one reading pretest (different from the Maryland assessment or the alternative assessment); the pretest scores are repeated with each measure for ease of reference.

[b] The "1993 adjusted means" are the 1993 mean scores statistically "adjusted" for the September 1992 pretest scores.

measures, third graders in the control schools have traditionally outperformed third graders in the participating schools. This was apparent in five years of CTBS data and on the 1992 baseline measure in mathematics. Therefore, one way of interpreting the between-year and covariance analyses together is to say that the assessment project helped participating students "catch up" to the control students in math achievement. From all indications, this would not have occurred without the project.

We also noted qualitative changes in students' answers to math problems which suggest that at least in some project classrooms whole groups of students were having opportunities to develop their mathematical understandings that had not occurred previously. Figure 3 and Tables 8 and 9 were constructed to provide a qualitative summary of student responses to a sample problem and to illustrate what slight improvements in student scores mean substantively. The two classrooms that showed the greatest gains from 1992 to 1993 in the low socioeconomic participating and control schools were selected for comparison (Table 8). Both teachers' classrooms showed an effect-size gain of .27 from 1992 to 1993 on the Maryland Mathematics Assessment. However, for this particular problem we think we see a greater gain for the participating classroom—one that suggests that a whole classroom of typically poorly performing students had developed knowledge of patterns and mathematical tables that this teacher's students had not understood the previous year. At the top of the scale there are no more right answers in 1993 than in 1992. However, in 1993 84% of the children in the participating classroom could complete the table (Categories I-V), whereas in 1992 only 34% of the same teacher's students could complete this part of the problem. The percentage of students in the participating classroom, who could write explanations describing a mathematical pattern or telling how they used the table (Categories I, III, or IV), also increased substantially, from 13% to 55%. Even students who took the *wrong* answer from their table could describe the pattern:

- I counted by fours which is 60 the(n) I went in the ones which is 15.
- I counted by 4 and ones and came to 60.
- First I went up to 15 pitchers. Then I made 60 cups.
- (60) first I cont'd by one's then I contid by fors
- first I saw that the where counting by 4s So I counted by fours. until there was no rome and got the answer 57.
- (15) I counted by 4s and I lookt at the top one

Figure 3
Sample Student Responses on Maryland Mathematics Assessment Problem Set Two
(Lemonade Step 4) Illustrating Key Qualitative Categories

I. Extends Table, Answers correctly, Explains either pattern or point in chart.

STEP 4 Now you want to know how many pitchers you will need for 46 cups of lemonade. You can see from the table below that a one-quart pitcher will hold 4 cups, and 2 one-quart pitchers will hold 8 cups. Continue the pattern in both rows of the table until you find the number of pitchers needed to hold 46 cups of lemonade.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

12

Explain how you got your answer. Write on the lines below.

I looked at the patteren and saw that there was not a 46. so I took 48 so there is also some for my friend and I.

Explain how you got your answer. Write on the lines below.

From pichers #11 to 12 it whent 44 48 cups so I just put 11½

Figure 3 (continued)

IV. Extends table, Wrong Answer (60, 15, 11, other), Explanation describes pattern.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

15

Explain how you got your answer. Write on the lines below.

I counted by fours which is 60 the I went in the ones which is 15.

Explain how you got your answer. Write on the lines below.

On the cups as you go along you count four more each time

Explain how you got your answer. Write on the lines below.

first I sow that the where counting by 4s So I counted by fours. until there was no rom and got the answer 53.

Table 8

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) From the Classrooms With the Greatest Gains in Low Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Extends table, Answers correctly, Explains (explains either pattern or point in chart).	13%	13%	31%	19%
II.	Extends table, Answers correctly, Inadequate explanation.	4%	0	8%	12%
III.	No answer but stops table at right place, Explanation describes pattern.	0	0	0	0
IV.	Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	0	42%	8%	4%
V.	Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	17%	29%	8%	35%
VI.	Cannot extend table.	63%	8%	46%	31%
VII.	Blank.	4%	8%	0	0

Table 9

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) From the Classrooms With the Greatest Gains in High Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Extends table, Answers correctly, Explains (explains either pattern or point in chart).	19%	43%	56%	43%
II.	Extends table, Answers correctly, Inadequate explanation.	8%	0	0	4%
III.	No answer but stops table at right place, Explanation describes pattern.	0	5%	0	0
IV.	Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	12%	29%	39%	9%
V.	Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	31%	9%	0	30%
VI.	Cannot extend table.	31%	9%	6%	13%
VII.	Blank.	0	5%	0	0

In the control low SES classroom the percentage of students writing explanations actually declined from 39% to 23%. Obviously for the two teachers to have the same positive gain in total score, there must be other problems where the control class gained relatively more. However, the qualitative analysis suggests that the gains in the control classroom tended to be more randomly distributed, with individual children earning a few more points here and there, but with no systematic shifts like the one just described. We are more inclined to attribute changes of this type to changes in instruction.

In Table 9, 1992 versus 1993 comparison data are shown for the same problem but for the two "best" classes in the highest socioeconomic pair of schools. Note that selecting the best class in the control school meant selecting the class with the smallest decline (ES = -.20) on the Maryland Mathematics Assessment, given that all classrooms in this school started higher in 1992 than any other classrooms but declined slightly in 1993. In contrast, the best classroom in the matched participating school showed a substantial improvement (ES = .53) and caught up to where the best control classrooms had been the year before.

Although the level of student performance is much higher in Table 9 than in Table 8, corresponding to the difference in socioeconomic level of the schools, the participating classroom in Table 9 still shows specific improvements in student performance that can be associated with the project intervention. Of course, there are more right answers (Category I), 43% versus 19% in 1992. More importantly, however, in 1993 77% of the children in the participating classroom wrote mathematically adequate explanations (Category I, III, or IV) about how they solved the problem. This proportion is in contrast to 31% who wrote explanations in the same teacher's classroom the year before. In the control classroom, 95% wrote adequate explanations in 1992 but only 52% could do so in 1993. As explained previously, we are more inclined to attribute these declines to population changes rather than to a decline in the quality of teaching, especially because all classrooms in the control school were affected. Table 9 also illustrates the increased ability of students in participating schools to extend a mathematical pattern or complete a function table. In 1992 only 70% of the participating children could extend the table (Categories I-V), but this percentage increased to 86% in 1993 making the participating classroom more comparable to the high levels achieved in the control classroom both years (95% and 86%, respectively).

Samples of student responses to a different problem or portion of the lemonade task are presented in Figure 4. Again we have chosen to illustrate the qualitative categories where students wrote explanations; these answers received either whole or partial credit in the quantitative scoring. This problem was much more difficult for children across schools and did not show much of an improvement for the low socioeconomic best classroom. There were no more right answers than in 1992, but 27% of students wrote mathematically adequate descriptions of the pattern (Category V) compared to 0% in 1993. However, a similar improvement occurred in the low SES matched classroom.

Category V responses show some of the richness of the students' answers and also help us to understand why many students found this problem more difficult. In every classroom there were some students who could count by fours when they got to step 4 but had trouble with steps 1-2 because they extended the table downward without looking at the left-right correspondence. They were able to explain what they were thinking mathematically in a way, in fact, that revealed their misconception:

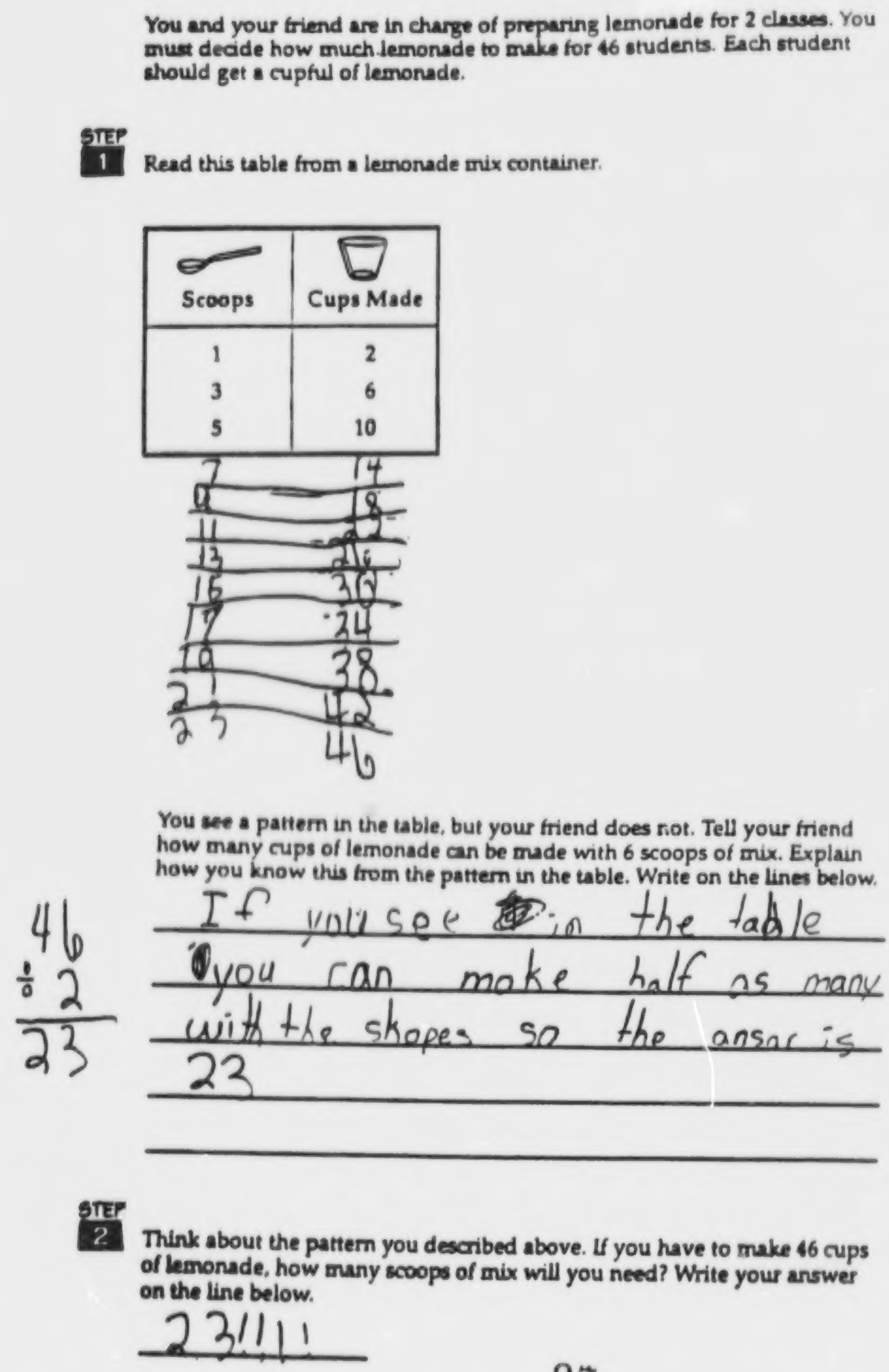
- Yes I do see a pattern, on the side with the spoon it counts by 2's were there's a cup it counts by fours.
- because on scoops it's go 1, 3, 5, I saw that their doing all odd so I put odd why cups was all even and 4 in the mitel. What I mean is 2 + 4 = 6 and 6 + 4 = 10 and so on.

The high SES classroom did show a substantial gain on this problem; the data for the comparison participating and control classroom are shown in Table 10. From 1992 to 1993 the percentage of students who wrote mathematical explanations (and extended the table) increased from 27% to 57% (Categories I, III, V). The corresponding change in the control classroom was from 45% to 35%.

The qualitative analyses of student answers on the Maryland Mathematics Assessment afforded us an appreciation of what an effect size of .13 means in terms of substantive increases in student learning. Significant shifts in specific categories occurred in participating classrooms but not in

Figure 4
Sample Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Steps 1-2) Illustrating Key Qualitative Categories

I. Right answers, Explanation describes pattern (includes minimal explanation 6 + 6 = 12).



You and your friend are in charge of preparing lemonade for 2 classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10
7	14
9	18
11	22
13	26
15	30
17	34
19	38
21	42
23	46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

$$\begin{array}{r}46\\ \div\, 2\\ \hline 23\end{array}$$

If you see in the table
you can make half as many
with the skopes so the ansar is
23

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23!!!!

Figure 4 (continued)

I. Right answers, Explanation describes pattern (includes minimal explanation 6 + 6 = 12).
Additional Examples

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

you add the same neber :
so exapel (6+6=12) (6X2=12)

If you put 1 scoop it will make
2. Then if you have 3 scoops it
will make 6. So every scoop you
do you will have to dubble
that number.

With 6 scoops of mix you
should be able to make 12 cups
of lemonade I figred this out
because 1+1=2, 3+3=6, 5+5=10
so 6+6=12 so that means you
have 12 full cups of lemonade.

$$\begin{array}{r} 23 \\ +\,23 \\ \hline 46 \end{array}$$

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

Figure 4 (continued)

V. Attempts to extend table but focuses on Left or Right column, not Left:Right pattern OR sees 1:2 pattern but can't apply to get answers.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10
7	14
9	18
11	22
13	26
15	30
17	34
19	38
21	42
23	46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

because on scoops it's go 1,3,5 I
saw that their doing all odd so I
put odd why cips was all
even and 4 in the mitel.
What I mean is 2 +4 =6 and 6+4=10
and so on.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

Table 10

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Steps 1–2) From the Classrooms With the Greatest Gains in High Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Right answers, Explanation describes pattern (includes minimal explanation, 6 + 6 = 12).	19%	24%	39%	9%
II.	Right answers, No explanation (but may show 23 + 23 = 46).	0	0	6%	9%
III.	Gets 12 cups with adequate explanation but cannot extend to 46 cups.	8%	9%	0	0
IV.	Gets 12 cups, Inadequate explanation, (wrong or no extension).	4%	0	0	4%
V.	Attempts to extend table but focuses on L or R column, not L/R pattern, OR 1:2 correspondence without answers, Explains thinking.	0	24%	6%	26%
VI.	Wrong answers, Explanation not based on chart or only restates answer.	58%	9%	33%	48%
VII.	Wrong answers, No explanation	4%	29%	0	0
VIII.	Blank.	8%	5%	17%	4%

control classrooms and are associated with the kinds of mathematical activities introduced as part of the project. In many cases this meant that students in the middle and bottom of the class were able to do things that their counterparts had not been able to do the previous year. These changes were confirmed quantitatively by significant shifts out of the lowest two quintiles for two of the three participating schools; (cut scores for quintiles were defined in 1992 on the total sample and held constant for 1993).

Conclusions

A fairly elaborate research design was implemented to evaluate the effect of a year-long performance assessment project on student learning. Maryland third-grade assessments in reading and mathematics and another alternative mathematics test served as independent measures of student achievement, separate from the classroom assessments developed as part of the project. End-of-year results in 1993 were compared to baseline administrations of the same measures in 1992 and to control school performance using analysis of covariance.

Results in reading showed no change or improvement attributable to the project. Third-graders in the participating schools did about the same on the Maryland Reading Assessment as third-graders had done the year before, and there were no significant differences between participating and control schools. In mathematics there were also no gains on the alternative assessment measure. However, small quantitative changes and even greater qualitative changes did occur on the Maryland Mathematics Assessment.

It is possible to offer both pessimistic and optimistic interpretations of the study results. Most significantly, from a negative perspective, it is clear that introducing performance measures did not produce immediate and automatic improvements in student learning. This finding should be sobering for advocates who look to changes in assessment as the primary lever for educational reform.

Of course, there were mitigating factors that help to explain and contextualize the lack of dramatic effects. First, we did not "teach to" the project outcome measures. For example, the classroom use of written summaries to assess meaning-making should have given students more experience with certain open-ended responses on the Maryland Reading Assessment. However, we did not introduce any other item formats from the outcome measure such as comparative charts or story webs. We should also note that the level of text difficulty in the Maryland assessment was quite high. In retrospect, we might have included additional, easier texts to be more sensitive to gains by below-grade-level readers.

Similarly, in mathematics we worked on explanations and used function tables as one of several problem-solving strategies (along with "guess and check," draw a picture, and "use cubes" [make a model]) but did not use formats that conformed specifically to the Maryland assessment. It is reasonable to assume that teachers might have behaved differently and imitated the outcome measures more closely if our 1992 baseline administration and anticipated 1993 measure had been imposed by an external agency for accountability purposes. Such practices could very likely have heightened the improvement of outcome scores, but then the question would arise as to whether the increased scores validly reflected improvement in students' understanding.

When we showed project teachers the outcome findings (in fall 1993), they were disappointed but offered an explanation regarding the "intervention" that jibes with our own sense of the project's evolution. Despite the level of workshop effort throughout 1992-93, by Christmas project "assignments" still had not been assimilated into regular instruction. Although we have evidence of changes beginning to be made in the spring term (Flexer et al., 1994), many teachers said that they did not "really" change until the next year (1993-94) (beyond the reach of the outcome measures). Several teachers argued that they did not fully understand and adopt project ideas and assessment strategies until they began planning and thinking about what and how to teach this year. This view is consistent with the literature on teacher change. Fundamental and conceptual change occurs slowly. Furthermore, changes in student understandings must necessarily come last, after changes in teacher thinking and changes in instruction.

We also note that the apparent gain in mathematics compared to zero gain in reading might have occurred because teachers had "further to go" in mathematics than in reading. If we take district curriculum frameworks as the standard, which are consistent with emerging professional standards in the respective disciplines, most teachers in the participating schools had already implemented some instructional strategies focused on meaning-making. In mathematics, the district frameworks were newer, and teachers were less familiar with them. Two teachers had tried out the Marilyn Burns (1991) multiplication unit the year before; but several more teachers decided to try it during the project year. Several were using manipulatives for the first time; several adopted materials to teach problem-solving strategies for the first time; and one group of teachers worked to develop new units in geometry and probability. Even when teachers did not understand them well or use materials optimally, these brand-new activities represented substantial shifts in the delivered curriculum.

In contrast to these apologies and caveats about why change did not occur, the cause for optimism comes from the small but real gains in mathematics.

Because of the project, most of the teachers in the participating schools spent class time on written explanations (especially what makes a good explanation) and on mathematical patterns and tables, which they had never done before. As a consequence, there were specific things that a large proportion of third graders in these classrooms could do on the outcome assessments, where before only the most able third graders had been able to intuit how to do them.

Our concluding advice is that reformers take seriously the current rhetoric about "delivery standards" and the need for sustained professional development to implement a thinking curriculum. Performance assessments—even with the diligent effort of most project teachers and the commitment of four university researchers—did not automatically improve student learning. The changes that did occur, however, confirm our beliefs that many more students can develop conceptual understandings presently exhibited by only the most able students—if only they are exposed to relevant problems and given the opportunity to learn. Performance assessments that embody important instructional goals are one way to invite instructional change, and assessments have the added advantage of providing valuable feedback about student learning. However, we would not claim that performance assessments are necessarily the most effective means to redirect instruction. When teachers' beliefs and classroom practices diverge from new conceptions of instruction, it may be more effective to provide staff development to address those beliefs and practices directly. Performance assessments are a key element in instructional reform, but they are not by themselves an easy cure-all.

References

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Champaign, IL: The Center for the Study of Reading, National Institute of Education, National Academy of Education.

Borko, H., Davinroy, K. H., Flory, M. D., & Hiebert, G. H. (1994). Teachers' knowledge and beliefs about summary as a component of reading. In R. Garner & P. A. Alexander (Eds.), *Beliefs about texts and instruction with text* (pp. 155-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burns, M. (1991). *Math by all means: Multiplication grade 3.* Sausalito, CA: The Math Solution Publications.

Davinroy, K. H., Bliem, C. L., & Mayfield, V. (1994, April). *"How does my teacher know what I know?": Third-graders' perceptions of math, reading, and assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davinroy, K. H., & Hiebert, E. H. (1993, December). *An examination of teachers' thinking about assessment of expository text.* Paper presented at the annual meeting of the National Research Conference, Charleston, SC.

Flexer, R. J. (1991, April). *Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Flexer, R. J., Borko, H., Cumbo, K., Marion, S., & Mayfield, V. (April, 1994). *How "messing about" with assessment affects instruction.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Hiebert, E. H. (1991, April). *Comparisons of student reading performance on standardized and alternative measures in high-stakes contexts.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12: Executive summary.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72*, 232-238.

Shepard, L. A., & Bliem, C. L. (1993,April). *Parent opinions about standardized tests, teachers' information and performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). *The role of testing in elementary schools.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

"HOW DOES MY TEACHER KNOW WHAT I KNOW?" THIRD GRADERS' PERCEPTIONS OF MATH, READING, AND ASSESSMENT[1]

Kathryn H. Davinroy, Carribeth L. Bliem, and Vicky Mayfield
CRESST/University of Colorado at Boulder

Historically, efforts for school reform and research on such reform have tended to focus on institutions, state agencies of education, local districts, school boards, teachers, administrators, parents, and test scores describing student achievement. Reform efforts are fueled by public opinion showing outrage at the failures of public schools and demanding standardized tests to measure academic achievement of students (Elam, Rose, & Gallup, 1992). Interest in the relationship between assessment and instruction has led some camps for school change to look to new assessment procedures as a route to improving the learning of U.S. children (Resnick & Resnick, 1992). Yet scarce research is conducted *with* school children to find out the effects of these assessment and instructional reform efforts on children's understanding of what it means to go to school and learn. The voices of those most directly affected by school change are frequently missing from the literature. Thus, education reform becomes mainly the business of everyone but the student.

If reform is to make a difference, it must gain access to the perceptions of those it seeks to educate. Outcome measures focused on student learning, even classroom-based performance assessments, give us access to only a portion of the successes and challenges of the reform picture. If current professional definitions of reading and math emphasize meaning-making, then one object of educational reform is to assure that children are acquiring these understandings in school. Some research has shown a relationship between children's understandings of subject matter and their approaches to the tasks of reading (Borko & Eisenhart, 1986) and doing math. These tasks may be part of instruction as well as assessment. Speaking with students about their perceptions of subject matter and assessment gives us another window through which to view reforms in assessment and instruction, as well as a view of changes in those student perspectives. Of what use is changing assessment and instruction to reflect new professional understandings of subject matter and learning if students continue to judge and to value their education in precisely those ways we seek to change? In addition to delving into student understandings, assessing our own progress in implementing reform may depend not only upon assessment measures that we created, but also upon the words and ideas of students.

---

[1] Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1994.

Students have taken center stage in more recent, constructivist approaches to learning. Rather than acting as receptacles for knowledge, learners are actively involved in constructing knowledge from past experiences, from social contexts, through interactions with an environment that encourages the exchange and growth of knowledge (Bruner, 1977; Vygotsky, 1978). In this view, student understandings of what they are trying to learn are critical in guiding their efforts. This view that places the learner at the center suggests an education and assessment system that looks different from traditional schooling. One aspect of this shift from tradition is the need for alternative assessment practices—practices that address the process of learning, practices that involve the student in actively demonstrating her or his construction of knowledge. Performance assessments, then, are one strand of school reform that recognizes the centrality of the learner. Speaking with students about their school experiences with instruction and assessment can help us know better if new perspectives on learning have taken hold in the classroom.

As alternative forms of assessment make their way into schools and research, close attention should be paid to the effects of assessment reform on all parties involved, especially the learners. Are assessment reforms helping students internalize contemporary professional definitions of what it means to read and do math in school and other contexts? The present study, drawn from a larger project in which teachers developed and implemented performance assessments in their classrooms, investigates the children's perceptions of what reading and math are and how they understand their teachers' knowledge of them as readers and mathematicians. We ask: With the introduction of reading and mathematics performance assessments into the classroom, what are students' understandings of what it means to be a reader and mathematician and of what it means to demonstrate those abilities?

Conceptual Framework: Assessment, Subject Matter, and Children

One question to ask is whether *how* one is assessed influences one's ideas and definitions of *what* is being assessed. How students are asked to use and demonstrate their knowledge in a classroom may help mold how they perceive knowledge and knowing. These relationships among instruction, learning, and testing are part of the growing body of research surrounding assessment reform. Some studies illustrate relationships between assessment and instruction that show negative effects of standardized testing on instruction (Shavelson, Baxter, & Pine, 1992; Shepard, 1991; Smith, 1991). Other research shows that assessment influences, even drives, instruction—especially in high-stakes testing contexts (Lomax, West, Harmon, Viator, & Madaus, 1992; Shepard & Cutts-Dougherty, 1991). These studies raise questions about the sort of knowledge that is being emphasized in instruction as a consequence of formal assessment.

In light of these findings, some want to make assessment less narrow and more "authentic" (Wiggins, 1989)—that is, linked to more realistic learning goals—by using performances of competence to assess students' understanding of subject matter:

> Authenticity is more than face validity (Does the task look [for example,] like a reading task?) or curricular validity (Is the task consistent with the manner in which it is presented in the current curriculum?). A literacy assessment task is authentic to the degree that it resembles the way literacy is used in real life. It is not enough to be consistent with the curriculum, which itself may be disconnected from real-life literacy. A slightly less rigorous version of . . . authenticity . . . would be this: An assessment task is authentic to the degree that it promotes the use of literacy in the ways students are expected to use it in genuine communication acts. (Garcia & Pearson, 1991, p. 271)

If assessment can be shown to influence instruction, then, for example, literacy assessments that reflect real-world conceptions of reading and writing might guide instruction toward more meaningful representations and toward a reformed, "thinking curriculum" (Resnick & Resnick, 1992). The same holds true in mathematics. Some educational reformers tout a relationship between assessment and instruction as an effective way of shifting classroom practices from a skill-based philosophy to higher order thinking, as well as a way of providing educators with a more complex and complete picture of the learner. Implementing performance assessments might be one way to both initiate and measure shifts in instruction. Concomitant shifts in student perceptions of reading and mathematics should also appear when instruction and assessment focus on thinking.

Subject area specialists in literacy and mathematics have supported the ideal of a thinking curriculum and of authentic assessment through various public statements of what constitutes reading and math and of what substantiates "excellence" in these areas. In *Becoming a Nation of Readers*, the Commission on Reading stated unequivocally: "The majority of scholars in the field now agree on the nature of reading: Reading is the process of constructing meaning from written text" (Anderson, Hiebert, Scott, & Wilkinson, 1985, p. 7). The Commission worried that when reading is taught and tested as decontextualized skills, students may very well learn and understand reading as discrete skills—not as an opportunity to make sense of text and the surrounding world. Reading as meaning-making has implications for instruction as well as assessment that include attention to strategies for fluency and extending comprehension into analytic and critical thinking. When reading is understood by teachers as meaning-making, both instructional methods and assessment change (Valencia, Hiebert, & Afflerbach, 1994). The effects of these changes on students' understanding of reading with the introduction of performance-based assessments are less well known.

In mathematics, the National Council of Teachers of Mathematics (NCTM) has formally adopted and published a set of curriculum and evaluation standards in order to ensure quality, indicate goals, and *promote change* in school mathematics (our emphasis, NCTM, 1989; NCTM, 1991). The standards document is built around a vision that includes learning to value mathematics, becoming confident in one's own ability, becoming a mathematical problem solver, learning to communicate mathematically, and learning to reason mathematically. The NCTM standards address not only what mathematics is and what it means to know and do mathematics, but also what teachers should do when they teach mathematics and what children should do when they learn mathematics. Significant changes in both instruction and assessment are required in order to implement the NCTM standards. Currently, researchers are studying what classrooms look like as teachers implement changes toward a meaning-centered approach in which students make conjectures, explain their reasoning, validate their assertions, and discuss and question their own thinking and the thinking of others (Ball, 1993; Lampert, 1990; Wood, Cobb, & Yackel, 1991). However, as was the case with reading, rarely do we hear how these reforms are perceived by the students.

Research on student perceptions of learning and subject matter has been embedded largely in studies of metacognition (Baker & Brown, 1984) and attitudes toward *learning to* read or do math. Studies concentrating on student views have generally emphasized *attitudes* about subject matter and self-assessments of ability (Stipek, 1981; Stodolsky, 1985; Sturtevant, 1991). Others have focused on students' conceptions of reading as related to their experiences learning to read (Borko & Eisenhart, 1986). And, while some research has looked at students' perceptions of math as subject matter (Cobb, 1986; Desforges & Cockburn, 1987; Schoenfeld, 1989), few researchers have related these perceptions to assessment.

Student ideas about reading, math, and assessment should be more closely examined as standards for subject matter competence are disseminated across regions, districts, and schools. Calls for changes in definitions and standards related to mathematics and reading and parallel changes in expectations for learners have made educators aware of the need to reconceive their instruction and assessment practices. Outcome measures give us only one view of these changes. Speaking with students may give us a more complete view. Researchers, as scribes and interpreters of educational reform, must look to multiple sources of information, not just test scores and teacher reports of classroom practice, as markers of reform progress. We must hear more frequently from the students. What are their perceptions about the relationship between what they do in their classes and the methods their teachers use to determine their progress? These questions are valid even in classrooms where the change process is only beginning. This paper attempts to get at student understandings of these concepts by asking:

- What are students' perceptions of math and reading?
- What are students' perceptions of their teachers' assessment practices?

    What do students recognize as modes of assessment: reading aloud, author's chair, chapter tests, quizzes?

    What do students see as criteria for assessment: behavior, percent correct, effort, level of difficulty?

- What connections do students make among their definitions of reading and math, instruction, and assessment?

Methods: The Larger Project and a Context for Understanding the Children

We interviewed third-grade students from the classrooms of the three schools whose teachers took part in the larger project. The participating schools are located in working-class and lower-to-middle-class neighborhoods on the outskirts of a western metropolitan area. The district for this research was selected on three criteria: (a) an ethnically diverse student population, (b) a history of standardized accountability testing, and (c) district willingness to seek a two-year waiver from standardized tests in the three schools selected to participate. As principals and third-grade teachers volunteered to participate, it is possible that the communities in this project were predisposed toward school reforms efforts such as performance assessments (Shepard & Bliem, 1993). During the 1992-1993 school year that we worked with these teachers, they were exempt from the pressures of preparing for districtwide standardized tests. Rather, they spent the school year developing and implementing performance assessments.

The project began with each school-team of teachers choosing a reading and mathematics instructional goal for the semester that could be measured using performance assessments. With the aid of literacy and mathematics experts, the teachers agreed upon goals that they perceived as important for third graders to attain, and which were in keeping with district curriculum guidelines. The research team was then able to provide guidance and materials as the teachers implemented performance assessments aimed at measuring student progress toward those goals. In reading, the teachers chose to focus on fluency and comprehension with narrative text as their goals for the first semester. The literacy expert suggested and the teachers incorporated running records and summaries as assessment tools compatible with these goals. In the spring, as teachers included expository text in their reading instruction, both of these assessment tools were again used to gather data about fluency and understanding. Comprehension, as demonstrated with the written summary, was evaluated using a 4-point scoring rubric which the teachers negotiated during workshop sessions (see Appendix A). As school-teams, they established the criteria for each score and developed the language that would frame those criteria. They also outlined instructional activities that would introduce and illustrate the criteria for students. One teacher even reprinted the scoring rubric and attached it to every student-written summary as a way to remind students of the standards for an excellent summary. Thus, summaries became a primary mode of instruction and assessment during the winter and spring months.

In math, the teachers elected to focus on the students' communication of their mathematical understanding. This goal took the form of explanations with the students being asked to explain the thinking that had led them to a solution of a math problem. This instructional goal necessitated the use of richer problem-solving tasks—tasks often provided by the mathematics specialist. As in reading, some teachers elaborated on explanations to a greater extent than others. One teacher wrote up a rubric similar to the summary rubric in reading as a way to capture both correct answers and good explanations. Others chose to grade responses separately along these two dimensions. In all classrooms, explanation of math problems was included in the mathematics instruction to some degree during our work with the teachers.

Because the entire project and treatment were structured according to subject matter, when we set out to talk with students, we devised an interview protocol that asked questions about mathematics and reading separately (see Appendix B). The interviews were counterbalanced so that one half began with reading questions and the other half with mathematics. For this study, we interviewed two students from each of the 13 classrooms at three times during the school year (fall, winter, and spring). Initial interviews were conducted in September 1992, with 26 students. We obtained a random sample of students, stratified according to teacher judgment of achievement, with 9 students categorized as high achieving, 8 students in the middle group, and 9 judged to be low achieving. The teacher judgments were revised for the winter sample, as the teachers thought they knew their students better by then. The second round took place in February 1993, with 20 new students and 6 repeats chosen randomly from the fall sample. The final round of interviewing occurred in May 1993, with again, 20 new students and the same 6 repeats. We originally designed the study to interview the 6 repeating students each time in an attempt to assure that "student" was not a factor of any change in student perspective that we might see.

Based on preliminary analysis of the data, we did not observe much change, so the 6 repeats were analyzed as part of the whole sample for each round of interviews. This resulted in a total of 78 interviews. Of these 78, one complete interview and one reading portion were lost due to technical failure, and another was omitted from the analysis due to a language barrier problem. This resulted in 75 interviews in reading and 76 in mathematics.

Our protocol included a number of probes enabling us to delve beyond single-word answers and to attempt to assure that students understood our questions. We conducted interviews during the regular school day in the hallways near students' classrooms. In a few instances, we interviewed students inside their classrooms while the rest of the class was busy away from the room. The interview protocol also included an invitation to students to retrieve items, papers, projects, and such—items that they felt their teachers used to determine how well they could read and do math. We sought to elicit students' perceptions of how their teachers understood their reading and math abilities and what evidence they believed their teachers consulted in order to form assessment judgments. However, given our knowledge of instructional and assessment changes that were going on in the classrooms, eventually we probed beyond what the students told us if they did not volunteer information related to the overall project. Thus, a subtle change in our protocol occurred between the fall and subsequent interview times: Upon recognizing the dearth of information from our fall interviews, we went to greater lengths to get the students talking in February and May by asking them specifically about summaries in reading and explanations in mathematics.

Analysis

Our analyses began with the authors transcribing all interviews. From these transcriptions, a careful reading of a sample drawn from the fall, winter, and spring interviews led us to a first-level coding system. The codes corresponded to our questions about understanding subject area, instruction, and assessment. We coded student responses as *definition* (of math or reading), as *instruction* (relating to their activities during the school day), and as *assessment* (their perception of their teachers' knowledge about them as learners). In describing our findings, student responses coded as "instruction" have been incorporated into discussions of definition and assessment because instruction tended to overlap heavily with both definition and assessment, and because these were our primary research categories. The three coding categories were further refined to differentiate between spontaneous responses and those requiring a greater degree of probing. We used this coding scheme systematically to categorize each student response to interview questions and examined the coded responses for patterns related to our research questions.

Findings

Our findings are organized according to subject matter following the categories of definition and assessment. We present definition first because we expected that students' concepts of reading and math would be tied to their perceptions of assessment. In the discussion that follows, we examine this expectation.

Reading Findings and Discussion

The reading portion of the interview began by asking students if they saw themselves as readers. We envisioned that their answers to this and to follow-up questions about what they read and why they read would help us understand their definitions of reading. In response to the question "Are you a reader?" most replied "yes" but a few lower achieving students said "no" or "not really" (see Table 1). This pattern remained constant throughout the school year.

Table 1

"Are you a reader?"

Response	Fall (*n* = 24)	Winter (*n* = 25)	Spring (*n* = 26)	Total (*N* = 75)
Yes	18	19	21	58
Sort of	1	4	2	7
Not really	3		1	4
No	1	1		2
Not asked	1	1	2	4

When asked why they read, nearly two-thirds of the students spoke of the enjoyment they receive from books. "Because it's fun to read . . ." "Because it's fun stuff. You can use your imagination and things" (see Table 2). These

Table 2

"Why do you read?"
Responses of Students Who Call Themselves Readers (*N* = 68)

Category	*N*
ENJOY It's fun. I like it.	42
LEARN To learn new things and get information.	27
INTEREST It's interesting.	9
WORDS I can learn how to spell new words and what they mean.	10
GROW UP I will need the skill as an adult.	4
OTHER	6
NOT ASKED	1

Note. Students often gave more than one response, so total will reflect more than the number of students interviewed.

responses about how fun reading is or how much they like it were often combined with comments about how they read to learn. In fact, nearly one-half of the students suggested that reading helped them by teaching them new things.

> Well, sometimes I read to learn, sometimes I read to entertain myself, and I just like to read a good book.

> Because it is sometimes interesting and it teaches you things. And sometimes it makes me laugh.

In general, these students see the act of reading as a useful venture, either to gather information or to amuse themselves. Indeed, they seem to recognize and appreciate the meaning-making involved in reading.

Even those who refrained from labeling themselves as "readers" appeared to interpret reading as a meaningful and useful activity:

> Because I'm slow, like, when you read, if you don't get a word and I mean, I like to read, but I can't read very well.

> . . . I have trouble with it. . . . I don't know any of the words.

And when asked what kept them from being readers, the students explained that their inability to read stemmed from word-level difficulties rather than from comprehension problems.

We sought to further understand the students' conceptions of reading by continuing with "I want to you think about what you do at school each day. When do you read?" We were hoping to hear about reading "period" as well as other instances of needing to understand text. Invariably students' initial response was to give us a time of day when reading instruction occurs in their classrooms: "At the end of the day," or "after Writer's Workshop." When probed, many students mentioned DEAR time, an acronym for "Drop-Everything-And-Read." They also spoke of reading at free time or whenever they completed their prescribed tasks. "We just, during reading time and sometimes when we have free time, like when we have finished our work and there's nothing to do, we read a book." But the end result was always the same: Regardless of when it happened, students read *books* and preferably chapter books. Only when explicitly probed did a few individuals admit to reading newspapers or magazines in the classroom.

Interestingly, a few lower rated students in the fall did talk spontaneously about having to read their math books. Story problems in math were, for this group, frequently mentioned when asked about reading at other times of the day. "Like instructions. Like to do a math paper, or something we don't know." In the spring, mid- and high-rated students also mentioned reading in other areas. "I have to read a math problem, I have to read what they ask you." In addition to subject-specific reading, a few children commented on other examples of reading during the school day. "We read our things that we put down, that we write. And we read stuff that she [the teacher] puts on the chalkboard."

We also asked students to reflect on any reading that they did outside of school, at night and on the weekends. As with school reading, the spontaneous reply was always books. When they read those books was usually at night or in the afternoons while playing school. When probed, some children would occasionally offer instances of reading the newspaper, magazines of interest, or comic books. Occasionally we heard about contexts other than book-reading:

> And sometimes when we're going for a trip down the road I like to read the signs. To see where we're going and stuff.
>
> Sometimes I like to read like words that are just on other things like on a cup I like to read that or something. Or like on this machine [the tape recorder] or something I'd like to.

As was true in school, it was rare to hear any students mention reading as meaning-making other than with books, like with travel signs or other environmental print.

Once we had some understanding of what children perceived reading to be, we turned to assessment issues. Our second set of questions was aimed at eliciting students' understanding of their teacher's assessment in reading. We began by posing the question "Does your teacher know how well you can read?" We continued by asking how she knew, what they did to help her know, or what she thought if they weren't sure. Most thought that their teacher did know, and the number who said "no" or "I'm not sure" decreased during the course of the year (see Table 3). However, they were often unable to articulate any modes or criteria of assessment. More than one child commented that "teachers just know," suggesting in their tone that the answer was obvious. "She just sort of, you know, knows if I'm doing a good job." "She just thinks I do." They had real faith that their teachers knew everything they needed to assign grades in reading because, well, *she is the teacher*.

A few students told us that they didn't know if their teachers could tell how well they read because the teacher never told them. Several commented:

> Not really. 'Cause I never hear her say it.
>
> Well, um, I think. I'm not sure. Um, she doesn't tell us how well we're reading or anything.

It seems that these students perceived that they did not receive much feedback, or perhaps they did not recognize it if they did.

When we imagined this study, we thought that by offering students the chance to show us ways in which their teacher knows how well they are doing they would lead us through their classroom, pointing to their work on walls, their portfolios, and the author's chair, for example. Unfortunately, it seems these children rarely see or have time to digest evaluated work. More typically, the work that is returned to students is sent directly home as evidence to the

Table 3

"Does your teacher know how well you can read?"

Response	Fall (n = 24)	Winter (n = 25)	Spring (n = 26)	Total (N = 75)
Yes	15	21	24	60
No	2	1		3
Don't know	5	3	2	10
Not asked	2			2

parents of schoolwork. More than once a student waved off the idea that he or she might show us his or her work. "I took it home" or "I threw it away" were the common messages.

However, the vast majority of students were not distressed by a lack of evidence about what their teacher thinks. They assumed that their teacher knows what she is doing as well as what the students are doing without requiring support of the assumption. When we posed follow-up questions about *how* she knows, they often seemed perplexed, both that we would ask the question and perhaps because they didn't have an answer. Evidently, children seem unconcerned about their role in the assessment process though they are aware that they receive report cards on a regular basis.

Still, we persevered, encouraging students to think about how their teacher knows about their reading. We asked them to reflect on what they as students do that lets their teacher know. This often allowed the students to talk about instructional activities that constitute their reading time (see Table 4). For every instance that a child offered, we responded by asking if they thought their teacher used information from that activity to know how well they could read. We probed further by asking how the teacher knew they had performed well on the activity, or what she was looking for if she graded it.

The most often-cited reading instructional activity that also helped a teacher know how well a student read was that students read aloud to their teachers. Most of the children suggested this, directly or otherwise. They recognized this as a measure of fluency and familiarity with words. Even the child who didn't think his teacher knew how well he could read had as evidence that she didn't listen to him read. " 'Cause they don't listen to me read." A few students claimed that expression was the indicator being measured when they read aloud: "She's looking for if I know when I should start a new sentence and when to slow down on my sentence . . . "

As the year (and the larger project) progressed, we heard the students talking about running records. Some, but not all, explained that this was how their teacher knew they could sound out big words. "She listens to us read and then she puts checks down for how good we can read and then the words that we have trouble with, she puts 'em down, too." By spring, most who had participated in a running record viewed it as a means of assessing their ability

Table 4

"How does your teacher know how well you can read?"

Response	Fall (*n* = 24)	Winter (*n* = 25)	Spring (*n* = 26)	Total (*N* = 75)
READ	15	18	17	50
RUNNING RECORD	0	5	7	12
QUESTIONS	14	14	7	35
WRITING	10	11	11	32
SUMMARIES	0	10	9	19
WATCH	6	4	3	13
GRADE	3	0	3	6
OTHER	5	9	11	25

READ	Student reads passages out loud for class or teacher.
RUNNING RECORD	Student reads a few pages from book while teacher listens.
QUESTIONS	Student answers questions about what she's read, either orally or on paper.
WRITING	Student composes written responses except summaries, including worksheets, answering questions, story maps, lit logs, diaries.
SUMMARIES	Student writes project summaries.
WATCH	Teacher watches students work to make sure they're behaving.
GRADE	Teacher knows because "I'm in the highest reading group."
OTHER	Includes talking with teacher, quantity of material read.

Note. Students often gave more than one response, so column totals will reflect more than the number of students interviewed at each time.

to figure out words. However, this was not because the teacher shared her notes from the record with them. One student said that even though her teacher didn't share the paper, she could see all the check marks on it:

> She, when we're reading this book about weather and she came around to my desk, and she was writing down all the words I didn't know. And most of them I did know. She put a check mark if I know the word. . . . Well, she doesn't show us, but you can see 'em while you're reading.

The degree to which students understood the criteria for the running record varied from one child to the next.

> She has like a blank piece of paper and she looks at you and she puts like little checks. I don't know why. Maybe if you get it right. She sometimes writes a word or something.

Few other assessment measures were mentioned spontaneously by students, especially at the beginning of the year. For this reason, we asked students if they did any writing about what they read. They talked about doing worksheets in reading time, usually skills-based lessons on contractions or spelling, or answering teacher-directed questions about what they had read. Later in the year, they included summaries and story maps in this list of written products. While they described the questions and summaries as tasks to tell "what the story was about," they often viewed the assessment of these same products as primarily based on handwriting and punctuation in addition to, or sometimes instead of, comprehension.

> They like check your handwriting and check and see if it's right and see if you spelled it correctly.

> If they really have a lot and the whole paper's filled and things. If we fill out the whole paper, if we have question marks and things and we have the exciting part of the story.

By the end of the year, many high-rated students (and some others as well) could recapitulate the scoring rubric developed by the teachers in project workshops as criteria for good summaries. Of course, the rubric was interpreted into the students' language. They talked about getting the questions right or wrong, based on what was really in the story. The words "understand" or "comprehend" were not part of their vocabulary when referring to the rubric:

> Well, if you would get a Thorough if you got like all good parts but not too much and then it's interesting or not. If you got a Solid, you would get, you don't have all the parts and it's a little interesting. And a Some you would get if you don't have very many parts and it's not interesting, and a Little is like you have no parts and you just didn't know.

Another child who outlined her understanding of the rubric proceeded to show the interviewer examples of her work. When the interviewer asked what she needed to do to make her summaries graded "3s" into "4s," the girl replied, "You need to have perfect spelling, periods go where they're supposed to."

A theme that ran throughout discussion of different modes of assessment (reading aloud, writing, teacher watching) was that behavior played a major role as a criterion. Students thought that *any* classroom reading activity could be judged upon how well they behaved while they were engaged. Obviously, classroom management is an issue for classes averaging 25 in number. Four years of schooling have led these students to conclude that their behavior counts no matter what the task. "She watches us read, and if we're not reading we got our name on the board. And you have to make up that time at recess."

Clearly, from their responses, these students generally recognize reading as a meaning-making task. They cite learning and enjoyment as reasons for reading. They even appreciate that tasks at school call on their abilities to comprehend what they have read. Yet, this message gets confused when they are formally assessed on their reading ability. Without specific criteria for meaning-making and practice with those criteria, students believe that assessment activities are often aimed at measuring their handwriting, at punctuation, at word recognition when reading out loud. The idea of text as meaning seems to get lost.

As discussed above, students rarely saw instances of their work with grades and comments on it. Until a child receives her paper with a score on it and explanation for that score, she may not internalize the rubric and apply it to her own work. During the course of the project, students did become familiar and sometimes comfortable with the scoring rubric for summary writing. It seemed the rubric was most internalized by higher rated students or by students who helped negotiate the rubric into their language.

Math Findings and Discussion

In turning to mathematics, we again tried to determine students' definitions of mathematics before asking questions about assessment. We probed for their definitions by asking them first about when, during the day, they used math, then whether they thought they were somebody who could do math, what were some of the things they could do, why people needed to learn those things, and finally when they use math outside the classroom. Though we hoped for spontaneous responses that would include occasions in addition to "during math time" as a response to when they use math, they only admitted that math was used during other activities when we probed. When asked directly if they thought taking attendance or counting hot versus cold lunches involved math, several reluctantly agreed.

When we asked whether they thought they were somebody who could do math, almost everyone said they could. However, there were a couple of students (low to middle range) during each round of interviews who said "No," "I don't know," or "Sort of" (see Table 5).

We followed up this question with "What are some of the things you can do?" and had an overwhelming response pattern of a list of computational skills: "I can add, and take away and count"; and "like subtract. I don't know how to do times." In the spring this list grew longer with responses like "Well I can add, I can subtract, I can multiply, divide." Though the list of things to do in math expanded, it was still a list revolving around computation. In an effort to probe for a definition that went beyond computation, we questioned students about daily instructional time asking what they *did* during "math time." Some students revealed that during math time they occasionally played games, engaged in activities, and solved problems in addition to the worksheets and pages of math book problems that most students cited. However, they did not add these math games or problem-solving activities that they said they were doing in their classrooms to the lists of their abilities in math.

Table 5

"Are you someone who can do math?"

Response	Fall ($n = 25$)	Winter ($n = 25$)	Spring ($n = 26$)	Total ($N = 76$)
Yes	20	18	22	60
Sort of	1	3	1	5
No	1	1	1	3
Don't know	1		1	2
Not asked	2	3	1	6

Further into the interview we shifted the focus from their ability in math to the possible meanings of math for others. In general, these third-grade students felt people needed to learn computational skills "so we won't be stupid when we grow up and we won't learn nothing 'cause we won't have a job then," and because "it would help them be a better person." The occasional student was explicit about how people would use this type of math in a future activity. "So that they can balance their checkbook," or "When they get older they don't have to worry about counting on their fingers and if they give you a price you won't, like if you're at a bank, kinda like my mom is, you could add the money that the people need or want." Finally, they gave only examples of computation as uses of math outside of school if they gave anything other than homework or playing school. The students seem to have bought into the "an education will make you a better person" message, but, though they see math as useful, it is defined as computation.

The heart of our protocol tried to determine what were students' perceptions of assessment. We asked whether they thought their teachers knew how well they could do math, how the teacher knew, and what her criteria were. Most students were quick to respond that yes, their teacher knows how well they can do math. There were a few that said "I don't know" and only a couple "I don't think so" answers (one student qualified this response as being because it was still early in the year) (see Table 6).

The question of "how does she know" was not an easy one for the students to answer and sometimes got confused with how the student knew the teacher knew. For example, we heard responses such as "she says I'm one of the smartest students in the class" or "I'm in one of the high math groups." However, in getting further into the question, a few students mentioned things

Table 6

"Does your teacher know how well you can do math?"

Response	Fall (*n* = 25)	Winter (*n* = 25)	Spring (*n* = 26)	Total (*N* = 76)
Yes	19	22	22	63
No	3			3
Don't know	3	3	4	10

like "my teacher from last year told her" or "my mom told her," and most said something like " 'cause she grades my five minute tests" or "well, 'cause she looks at my papers." Most students seemed to have decided for themselves that their teachers use their tests and papers in math in order to determine how well they can do math (see Table 7). These tests and papers focused mainly on computation which seems to have helped develop their perceptions of mathematics as computation.

When we asked "What is she looking for on your papers?" or "How does she determine if you've done a good job?" the main criterion the students reported was right answers:

> Right answers, wrong answers, um, that's about it.
>
> If I get the answer right, she knows I was thinking hard.
>
> How much problems we get right and how much we don't.

Table 7

"How does your teacher know how well you can do math?"

Response	Fall (n = 25)	Winter (n = 25)	Spring (n = 26)	Total (N = 76)
PAPERS	13	18	19	50
TESTS	4	10	12	26
WATCHES	5	9	3	17
GRADES	2	1	0	3
OTHER	3	1	2	6

PAPERS	Math book problems and worksheets.
TESTS	Timed tests as well as other math tests.
WATCHES	The teacher watches the students as they work on their math, play games, do activities.
GRADES	Report card grades.
OTHER	Records from other schools/grades, mother told the teacher, student told the teacher, conferences.

Note. Students often gave more than one response, so column totals will reflect more than the number of students interviewed at each time.

Several combined the "right answer" criterion with handwriting, neatness, spelling, correct format, and behavior:

> That the answers are right and the handwriting's right and you borrow a lot.
>
> She looks for if the answers are right, neat handwriting and if you have space in it. Like space between this way and down.
>
> . . . if all the words were spelled correctly and stuff.
>
> She's watching for somebody paying attention so they can do their work. And nobody being so stupid and funny.
>
> Oh. I think she thinks I've done very good because I'm quiet and I do my math and I get it done right on time.

Some students indicated that effort was yet another criterion for doing a good job: "Because she saw me working hard." In general, these students felt their teachers knew how good they were in math by assessing for "right" answers and by watching for logistics and behavior. Even when probed, no students mentioned their teacher checking for understanding.

Of these papers, the assessment tool that seemed to make the biggest impression on the students was the timed math-fact test. When we asked what does a math test look like, one student responded: "It has a hundred problems on it and you have to get as many problems as you can down in five minutes." The timed test also seemed to help create the impression that math is something you have to do quickly in order to be good at it. Several students made comments like: "And I'm doing really good on my times test. I do them really fast. I'm getting really fast at them," or "Well, because I'm usually the first one done; I'm really good at it . . . I can get the number in my mind and just subtract or add it or multiply it real fast." The number of students who mentioned timed tests increased dramatically from 4 in the fall to 10 in the winter and 11 in the spring. Some teachers gave a lot of attention to timed tests according to the students: ". . . she looks at them and she says, well, we give people, we clap for the students that really improve by at least 10 or 15."

The larger project encouraged teachers to focus on creating assessments outside of the timed tests—a focus that shifted attention from computation to problem solving. As with summaries in reading, we worked with the teachers to create a performance assessment and rubric that asked students to write explanations of how they solved math problems. Because we had knowledge of the broader project and because we hoped students were moving beyond computation, we probed students to tell us if their teachers ever asked them to explain their answers. In this question we did see a shift between the fall and the winter/spring interview times (see Table 8). In the later interviews, more students responded that they did explain their answers to math problems, at least occasionally.

There were some students who when probed could also talk about the scoring rubric for an explanation of a math problem.

I: OK. How do you know if you've given a good explanation?

S: Cause she'll usually give us a number like 4 or 3 or 2 or 1.

I: Do you know what those mean?

S: 1 means you did it a little bad, and 2 is kinda good, and 3 is good, and 4 is good.

I: Do you know what you have to do to get a 4?

S: You have to get the answer right and you have to explain it good.

Table 8

"Does your teacher ever ask you to explain your answer?"

Response	Fall (n = 25)	Winter (n = 25)	Spring (n = 26)	Total (N = 76)
Yes	8	17	12	37
Sometimes	6	5	7	18
No	8	2	3	13
Not asked	3	1	4	8

These students also seemed to understand the contrast between the answer and the explanation of their solution. In reflecting on a 2 a student received on one paper, she said it meant "that I wasn't, that I didn't tell a lot about it, and I didn't have the right answer." However, sometimes what a student meant by explain was not always what we meant (nor what the larger project meant).

> I: Does she ask you to explain your answers ever?
>
> S: Yeah, sometimes.
>
> I: What does she ask?
>
> S: Like, uh, what's this answer, and I tell it to her and she says good and then I'm done.

It did not appear that the students understood that the teacher might be using these explanations as indications of how well they were doing in math. Though explanations were being scored, the value of the explanations as a *mathematical* activity didn't seem to influence student understandings of how their teachers knew they could do math.

Our data suggest a picture of third-grade students who view math as arithmetic problems that have right and wrong answers. Students say their teachers assess them in ways that are consistent with this view by giving them tests and problems to do and by grading them mainly for correctness. Though the third-grade curriculum is heavy in arithmetic skills, reformers would hope that these skills could be combined with equal weight with problem-solving abilities. There is significant emphasis on getting third graders to "know their facts." However, the NCTM standards are trying to get away from this emphasis especially in the decontextualized timed-test setting.

Some Cross-Subject Similarities

Our findings about student perceptions regarding reading, mathematics and assessment support contentions that reform takes time if perceptions and understandings are going to change significantly (Anders & Richardson, 1992; Borko & Putnam, in press; Richardson, 1992). Interestingly, in addition to subject-specific findings, issues of communication and student access to graded work cut across subject areas.

Communication with third graders proved to be an occasional problem across subject matter. These problems seemed to be associated with our inability to ask a question that students understood in the way that we did. Communication difficulties were signaled by lengthy periods in transcripts in which student and interviewer engaged in a rapid back and forth exchange that didn't seem to arrive closer to a definitive response. When 20 to 40 lines of transcript are devoted to phrasing and rephrasing of the interview question, clearly a miscommunication is occurring.

One example of a communication problem exhibited in the reading interview occurred when we probed with the question "Does your teacher ever ask you questions about what you're reading?" We, as interviewers, thought we might hear instances of students retelling stories and telling us this demonstrated their comprehension. We imagined students would respond with questions like "What happened to Bobby?" or "How was the problem solved?" But, while our third graders agreed that, yes, their teachers did ask them questions about their books, even with persistent probing students responded with examples most often in the form of "Do you like the book?" or "Is it the right level for you?" While these may be useful instructional questions, they did not tell us much about assessment, particularly assessment of understanding. Our parallel question in math, "Does your teacher ever ask you to explain your answer?" showed some similar though less consistent difficulties. When we probed students to talk about explaining problems, often they told us, "I tell it to her and I'm done." The differences between our expectations for responses and the actual student responses suggest either we weren't communicating with the children, or teachers actually did not ask the sorts of questions that we thought they were.

Another unexpected communication misunderstanding recurred when we probed students about how their teachers knew their reading abilities. Several answered by outlining how they knew their teachers knew they could read. The classic example was of the little girl who declared that her teacher knew she was an excellent reader "because I'm in the highest reading group." When probed about how the teacher knew to put her in this group, the student replied that it was because she was, in fact, a good reader. Students struggled with this question in math as well. When students tried to describe how their teachers knew how well they could do math, it was shifted or translated to how the *students knew* their teacher knew. Rather than describing a *process* by which teachers would form judgments, students were telling us how they knew their teachers' opinions, and this was based mainly on products or *results* of teacher assessment: I'm in the highest group; because I'm a good reader or math person.

Clearly at times, our students seemed to be trying to appease the interviewers. When the questioning went on long enough, they would list every instructional activity they could think of as possible vehicles for assessment. Literature logs, double entry diaries, the number of pages or books read were all offered as ways their teachers know how well they can read. Several explanations for the outcome are possible. The laundry list response could be an artifact of our interviewing style—students chose to list everything they do as a way to move beyond the question. Alternatively, the children could believe that such tasks—though very frequently ungraded—are used by their teachers to know how well they can read. In fact, talking with the teachers about their assessment practices, several mentioned the "gut feeling" they have about their students just from observing them every day (Borko, Flory, & Cumbo, 1993).

The research literature on student thinking and perceptions prepared us somewhat for the eventuality of communication problems with third graders. Indeed, the difficulties of interviewing young children may contribute to the paucity of studies about their perceptions and understandings of subject matter as related to instruction and assessment. Research on young students' thinking and perceptions may well be hampered by the fact that children have not yet developed a causal understanding of the relationship between ability and achievement. According to Wittrock (1986), "children's concepts of the causes of their successes and failures develop from a relatively undifferentiated state to a more analytic conception of the relations among ability, effort, and achievement . . . at about 7 to 8 years of age they distinguish these concepts from one another, and causally relate effort, but not ability, to achievement" (p. 304). Asking students if their teachers know how well they can read and do math assumes that students understand that teachers are concerned with achievement and go through an explicit process for judging student performance. It is possible that scholastic achievement may be a nascent concept for third graders. But this possibility should not deter others from pursuing students' ideas of school subjects. Only through open-ended conversations can we more deeply probe and understand whether students see reading as meaning-making and mathematics as broader than computation.

Another cross-subject phenomenon involved our repeated efforts to get students to share work they felt their teachers used to assess them; we were often met by blank looks or responses that indicated they weren't sure what we meant. Very rarely did our question about graded work that their teacher would use to determine how well they could read or do math result in physical artifacts—like a scored summary or math explanation. The majority of students did not show us any work. One inference that might be drawn from this cross-subject similarity is that students do not seem to get to see enough of their work after it is graded to understand the assessment value that it has for their teachers, nor to understand how classroom performance gets translated into teacher assessment. Papers get thrown away, filed away, or sent home in folders to parents. Typically the path of student work moved from student to teacher to parents or to a file, and rarely was work returned to students for individual and/or class discussion. One teacher told us that in an ideal world, papers would be graded and returned to students quickly, and would be followed up with discussion because students need the feedback in a timely way. Time-in-the-day continues to act as a major constraint in trying to vary the path of assessment artifacts and results.

Some Closing Thoughts and Implications

Our recommendations from this study begin with where we started: students' conceptions of reading and mathematics. Clearly, from their responses, these students recognize reading as a meaning-making task. They cite learning and enjoyment as reasons for reading. They even appreciate that instructional tasks at school call on their abilities to comprehend what they've read. Yet, this recognition gets distorted when they are assessed on their reading ability. These students believe that assessment activities are often aimed at measuring their handwriting, punctuation, expression when reading out loud. The idea of reading as meaning shifts to reading as word identification. Responses to mathematics questions took a slightly different tack. Whereas students described reading as meaning-making and reading assessment as skills-based, in mathematics they demonstrated consistency across definition of math and assessment of math. In both categories, math is arithmetic problems that have right and wrong answers. Teachers know how well students can do math as a function of how many right and wrong answers they produce—not according to their thoughtfulness in problem solving.

With respect to project-specific reading and math tasks, students were aware of these tasks though few saw them as ways their teachers were assessing them. In reading, teachers purportedly used summary writing to assess comprehension, but students maintained a skill-based stance by explaining that their teachers looked for spelling and other mechanics in assessing summaries rather than providing the "gist" of a passage as called for by literacy experts (Palincsar & David, 1991; Pearson & Fielding, 1991). Our student perceptions of summary writing align with findings of some of the teachers' understandings of summary as reported in an early project study. Findings from the fall semester in one school, suggest that teachers agreed to use summaries to assess comprehension but found themselves focusing on using summary as an end in itself (Borko, Davinroy, Flory, & Hiebert, 1994).

In math, the larger project introduced into classrooms a number of math activities with the intent that they would be used for both instructional and assessment purposes. While students occasionally described "explaining" their answers as one way their teachers knew how well they could solve a problem, for the most part students saw these math activities as opportunities to produce neat work and to work hard. Student definition and understanding of assessment continue to follow typical ways of perceiving math—as computation. Getting away from math being simply computation is one of goals of the NCTM standards, and some say that students will learn what is valued by what is assessed. If the students don't recognize that their teachers assess them on their understanding of mathematics, how can they come to value understanding? If we want students to better understand and value the meaning-making and communication involved in mathematics, we need to find ways to make these connections more explicit for the students.

With respect to reading, it seems teachers need to continue to work toward authentic performance assessments that extend and build on students' existing understandings of reading. In math, the task is more challenging. That students continue to perceive of math as arithmetic problems and that teachers continue to support this perception with assessments that align with that view reflect the problems in implementing the fundamental changes called for by NCTM—to shift the focus from skills to mathematical thinking.

As discussed above, in both reading and math students need to see instances of their work with grades and comments on it. Too often, it seems, the papers are handed in, never to be seen again unless it's in the folder for parents. And it doesn't appear to be enough to work together as a class to derive scoring criteria. Until a child receives her paper with a score on it and an explanation for and discussion of that score, she may not internalize the rubric and apply it to her own work. Our findings raise some interesting questions about how students relate their concepts of subject matter to assessment practices: Should they be made aware of every instance that will be scored? Would this awareness, like public scoring criteria, help them, especially low-achieving students, perform more to their potential? How can we help to initiate a culture in the classroom where students realize that everything they do helps their teacher know how well they can read and do math, so that their teacher can aid them in becoming better readers and mathematicians?

Reform efforts arise from every aspect of educational research. When assessment reform can help spur reform in instruction the many facets of education may be able to work together for general and lasting reform. But it is also the learners who must speak to these reforms. As efforts to demonstrate new and better instruction accelerate, students must be consulted more frequently as a gauge for reform success.

This paper is one beginning in trying to hear from the students about educational reform efforts, about how they understand their experiences in school, about how they form understandings of what it means to be a reader and mathematician. If it can be shown that performance assessments as an alternative to standardized and limited-format testing are effective in measuring student achievement; if it can be shown that instruction developed from performance assessments involves the whole classroom community in learning and assessing; and if it can be shown that students exposed to these new ways of assessing and instructing are constructing new meanings of subject matter—meanings that break with traditional and limited modes of understanding what it means to be literate and do math; then perhaps calls for a thinking curriculum may be closer to being realized.

References

Anders, P., & Richardson, V. (1992). Teacher as game-show host, bookkeeper, or judge? Challenges, contradictions, and consequences of accountability. *Teachers College Record, 94*(2), 382-396.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Washington, DC: The National Institute of Education, U.S. Department of Education.

Baker, L., & Brown, A. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 353-394). New York: Longman.

Ball, D. L. (1993). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. In T. P. Carpenter, E. Fennema, & T. Romberg (Eds.), *Rational numbers: An integration of research* (pp. 157-196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Borko, H., Davinroy, K. H., Flory, M., & Hiebert, E. H. (1994). Teachers' knowledge and beliefs about summary as a component of reading. In R. Garner & P. Alexander (Eds.), *Beliefs about texts and instruction with text* (pp. 155-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Borko, H., & Eisenhart, M. (1986). Students' conceptions of reading and their reading experiences in school. *The Elementary School Journal, 86*(5), 589-611.

Borko, H., Flory, M., & Cumbo, K. (1993, April). *Teachers' ideas and practices about assessment and instruction.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Borko, H., & Putnam R.T. (in press). Learning to teach. In R. C. Calfee & D. C. Berliner (Eds.), *Handbook of educational psychology.*

Bruner, J. (1977). *The process of education.* Cambridge, MA: Harvard University Press.

Cobb, P. (1986). Contexts, goals, beliefs, and learning mathematics. *For the Learning of Mathematics, 6*(2), 2-9.

Desforges, C., & Cockburn, A. (1987). *Understanding the mathematics teacher: A study of practice in first schools.* London: The Falmer Press.

Elam, S. M., Rose, L. C., & Gallup, A. M. (1992). The 24th annual Gallup-Phi-Delta-Kappan Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 74*(1), 41-53.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Garcia, G. E., & Pearson, P. D. (1991). The role of assessment in a diverse society. In E. H. Hiebert (Ed.), *Literacy for a diverse society* (pp. 253-278). New York: Teachers College Press.

Hiebert, E. H. (Ed.). (1991). *Literacy for a diverse society: Perspectives, practices, and policies.* New York: Teachers College Press.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer. Mathematical knowing and teaching. *American Educational Research Journal, 27*(1), 29-63.

Lomax, R. G., West, M. H., Harmon, M. C., Viator, K. A., & Madaus, G. F. (1992). *The impact of mandated standardized testing on minority students.* Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

Palincsar, A. S., & David, Y. (1991). Promoting literacy through classroom dialogue. In E. H. Hiebert (Ed.), *Literacy for a diverse society. Perspectives, practices, and policies* (pp. 122-140). New York: Teachers College Press.

Pearson, P. D., & Fielding, L. (1991). Comprehension instruction. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 609-640). New York: Longman.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Richardson, V. (1992). the agenda-setting dilemma in a constructivist staff development process. *Teacher and Teacher Education, 8*, 287-300.

Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for Research in Mathematics Education, 20*, 238-255.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments. Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *72*, 232-238.

Shepard, L. A., & Bliem, C. L. (1993, April). *Parent opinions about standardized tests, teachers' information and performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Shepard, L. A., & Cutts-Dougherty, K. (1991, April). *Effects of high stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Smith, M. L. (1991). Put to the test: the effects of external testing on teachers. *Educational Researcher*, *20*(5), 8-11.

Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology*, *73*, 404-410.

Stodolsky, S. (1985). Telling math: Origins of math aversion and anxiety. *Educational Psychologist*, *20*(3), 125-133.

Sturtevant, E. G. (1991). Reading perceptions of urban second graders. In T. V. Rasinski (Ed.), *Reading is knowledge* (pp. 63-69). Pittsburgh, PA: College Reading Association.

Valencia, S. W., Hiebert, E. H., & Afflerbach, P. (Eds.). (1994). *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

Vygotsky, L. S. (1978). Interaction between learning and development. In L. S. Vygotsky, *Mind in society* (pp. 79-91). Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1992). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA: MIT Press.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership*, *46*, 41-46.

Wittrock, M. C. (1986). Students' thought pieces. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 297-314). New York: Macmillan.

Wood, T., Cobb, P., & Yackel, E. (1991). Change in teaching mathematics. A case study. *American Educational Research Journal*, *28*, 587-616.

APPENDIX A

Scoring Rubrics for Narrative and Expository Summaries

For Narrative text:	For Expository text:
4 (Thorough) — Includes the setting, the characters, and plot. Told in an interesting way.	4 (Thorough) — Organized. Includes main idea and some support. Written in student's own words.
3 (Solid) — May be missing some important parts of the story. May have too many details.	3 (Solid) — Not completely organized, but still flows. Includes main idea and some support. May have copied some phrases directly from the text.
2 (Some) — Doesn't have the most important parts. May have some wrong information. Not told in an interesting way.	2 (Some) — May not have main idea, but includes some details OR may have main idea but includes wrong information. May include wrong supporting details.
1 (Little) — Doesn't make sense. Not enough information is given.	1 (Little) — Includes a lot of incorrect information. Focus is on unrelated details or only on supporting ideas.

APPENDIX B

Student Interview Protocol

Reading/Math

1. **First I'd like to ask you some questions about reading. Are you a reader?**

If yes, what do you read?
Why do you read? If "because I like it", why do you like it?

If no, what keeps you from being a reader?

2. **I want you to think about what you do at school each day. When do you read?**

Probe (if child focuses on formal reading time): Do you read things at other times?
If yes, what do you read?
If no, probe by asking about other subject areas like "what about social studies?", and then offering examples such as science books, magazines, newspapers.

3. **Does Ms. ______ know how well you can read?**

If no, what does your teacher think? Why do you think she doesn't know how well you can read (or why is that)?

If yes, how does your teacher know?

Can you give me an example? You can go to your desk to get something, or take me there to see things. What kinds of things is she looking for?

For example, "Can you describe (that activity) for me?", "What does your teacher do/look at?", "Can you show me (this activity)?"

For conferences, What do you do? What does your teacher do? What does she write? Do you see/hear what she's looking for? Does your teacher ask questions? About what?

How does your teacher decide if you've done a good job (for each mode of assessment)?
Do you help your teacher decide if you've done a good job?

Probe: Look around the classroom. Is there anything else?

4. **Now think about what you do after school, at night, and on the weekends. Can you think of times when you read outside of school?**

Probe for "what do you read (for each time mentioned)?"

If examples focus on school/homework, probe by offering examples: reading to siblings, read TV Guide, etc.

5. **Now let's go back to school and talk about math. Think about your school day and tell me when you do math.**

What do you do during that time?

> Probe (if student focuses on math time only): Are there any other times during the day when you use math?
> If yes, what do you do?
> If no, probe by offering examples of lunch count, attendance, etc.

6. **Do you think you are someone who can do math?**

> If yes, what are some of the things that you can do?
> Given examples, ask "Why do people need to learn these things?"
>
> If no, what keeps you from doing math?

7. **Does Ms. ____ know how well you can do math?**

If no, what does your teacher think? Why do you think she doesn't know how well you do math (or why is that)?

If yes, how does your teacher know?

> Can you give me an example? You can go to your desk to get something, or take me there to see things. What kinds of things is she looking for?
>
> For example, "Can you describe (that activity) for me?", "What does your teacher do/look at?", "Can you show me (this activity)?"
>
> For papers/worksheets, what do the questions look like? What does the teacher do? Do you know what she's looking for? What? Does she ask you questions? About what?

How does your teacher decide if you've done a good job (for each activity)?

> Probe: Look around the classroom. Is there anything else?

8. **Now think about after school, at night, and on the weekends. Can you describe for me a time when you use math outside of school?**

> Probe for "What do you do (for each time mentioned)?"
>
> If examples focus on school/homework, probe by offering examples: allowance, pages read in book, buying things.

9. **Is there anything else you want to tell me about reading or math?**

HOW "MESSING ABOUT" WITH PERFORMANCE ASSESSMENT IN MATHEMATICS AFFECTS WHAT HAPPENS IN CLASSROOMS [1,2]

Roberta J. Flexer
CRESST/University of Colorado at Boulder

Introduction

This paper reviews a year's work with third-grade teachers who introduced performance assessments in the hope of improving both instruction and assessment in mathematics. Our interest in this effort, and the staff development program we designed, drew upon ideas central to current reform in mathematics education and educational measurement. Participating teachers tried out many changes in their instructional and assessment practices. By year-end, teachers had increased their use of hands-on and problem-based activities, extended the range of mathematical challenges they considered feasible to attempt with third graders, and incorporated performance tasks and observations to replace or supplement computational and chapter tests.

This report also examines teachers' beliefs related to assessment and instruction in mathematics as they experimented with new assessments in their classrooms. More specifically, we examine patterns of stability and change that resulted from teachers' year-long effort to incorporate performance assessments into their instructional programs.

The current reform in mathematics education can be described by three sets of standards produced by the National Council of Teachers of Mathematics

---

[1] Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1994.

[2] We thank Abraham S. Flexer for his support throughout the project and for his editing of this manuscript. We also thank Carribeth Bliem, Kathy Davinroy, and Maurene Flory for their many hours of work on the project, particularly the hours of sitting through meetings with teachers, transcribing tapes, and checking transcripts. We give special thanks also to Pam Geist, a visiting researcher, for her very valuable contributions to the teachers and to the research team.

We are particularly grateful to the teachers who worked so hard for this project and to their district administrators and personnel.

(NCTM): *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), *Professional Standards for Teaching Mathematics* (NCTM, 1991), and *Assessment Standards for School Mathematics—Working Draft* (NCTM, 1993). (These sets of standards will be referred to in the rest of this paper as the NCTM *Standards*.) These standards grew out of work done in the late 70s, reported in 1980 in an *Agenda for Action* (NCTM, 1980), that was a reaction to the Back to the Basics Movement of the 70s. The curriculum, assessment, and instruction proposed in these NCTM *Standards* emphasize mathematical thinking, reasoning, problem solving, and communication. Students are expected to understand the mathematics they do and to model and explain their work. The emphasis is no longer on memorization of facts and the mechanical following of procedures. Mathematics is supposed to be relevant and contextualized. The content of the curriculum is supposed to be broader than numeration and computation, and to involve, for example, topics in geometry, probability, and data analysis. Algebraic ideas are to be brought into the elementary schools, giving younger students powerful tools for attacking problems.

Concurrent with this reform in mathematics education, a reform movement is underway in the measurement community. Researchers are investigating the extent to which instruction is influenced by standardized tests (Romberg, Zarinnia, & Williams, 1989; Smith, 1991). The standardized tests, then and now, focus on recall of facts and definitions and demonstration of computational procedures; and many teachers appear to respond by narrowing instruction to what is on the tests and in a format compatible with the tests. Teachers state their sense of responsibility for "preparing" their students for such tests. Their position is often justified by the high stakes some districts place on having their students perform well (Shepard & Cutts-Dougherty, 1991). A prior study by this CRESST-CU research group showed that elementary students in a high-stakes district were able to produce scores on standardized tests that did not hold up when the students were given other tests of the same material (Flexer, 1991; Koretz, Linn, Dunbar, & Shepard, 1991). In addition, the more the format of an alternative task varied from the corresponding standardized-test task, the poorer was students' performance. From these studies it appears that standardized tests in high-stakes contexts are having a deleterious effect on what students are learning in mathematics.

The response of many teachers to these tests is to omit or limit instructional time on untested topics and to teach others at the lower levels of thinking that match the tests.

In the late 80s there was a convergence of writings by mathematics educators who encouraged the adoption of the new standards of curriculum, evaluation, and teaching, for example, *Everybody Counts* (Mathematical Sciences Education Board, 1989), on the one hand, and by researchers in the measurement community (e.g., Shepard, 1989; Wiggins, 1989) who argued that standardized tests were having a negative effect on instruction and curriculum and were inadequate for promoting higher order thinking on the other. Curriculum proposed by the NCTM *Standards* is incompatible with standardized tests, but because standardized tests were in place, they were affecting what and how teachers taught. One approach to bring about the hoped-for changes in curriculum and instruction proposed in the *Standards* was to develop state or national tests that are more compatible with the *Standards*. Several state and one national assessment project took this approach and developed tests that included performance assessment tasks, for instance, Maryland, Kentucky, Massachusetts, Maine, and the New Standards Project. If the new tests require broader thinking, reasoning, and problem solving, then teachers would have to teach in such a way that their students were ready for these kinds of tasks. Here at last was a way to change curriculum and instruction—by adopting an end-of-year test that requires a different kind of performance than the old standardized tests. Support for this "top-down" approach to change comes from Gipps' (1992) report that performance assessment (the UK's Standardized Achievement Tasks, SATs) can have positive effects on instruction. But there are also questions about the effects *any* externally imposed test, even if more authentic, will have on instruction, particularly concerns about narrowing the curriculum (Shepard, 1991).

Another approach to change is a "bottom-up" approach in which teachers are helped to change their assessment program in ways that comply with the *Standards*, and are further helped to change their instruction to align it with their assessment, and similarly with the *Standards*. This is the approach taken in the current study, and this paper is a report of the effects of third-grade teachers' work on performance assessment in mathematics on their beliefs and practices about curriculum, instruction, and assessment. It is an account of their struggles and successes during an academic year—and of the ways they changed what they thought was important to teach, how they taught, and how they assessed the performance of children.

In this study we are concerned about the teachers' beliefs and practices with respect to what they value in mathematical performance, what school mathematics should be, how children learn, and how they should teach. Both from our own work with teachers and from that of other researchers (Battista, 1994; Cobb, Wood, Yackel, & McNeal, 1992), it is clear that teachers' beliefs about how children learn mathematics and the nature of school mathematics will very much influence their beliefs and practice about instruction and assessment in mathematics (see Figure 1). We did not intend to confront directly teachers' beliefs but expected beliefs would shift through work on assessment practices and, as it turned out, on instruction practices. We believe that belief and practice can be causally related in both directions, and that it is not only the case that a change in belief causes a change in practice. A shift in practice may lead to a shift in belief which can lead to further shifts in practice (see Figure 2). We know from the literature on teacher change (Borko & Putnam, in press; Nelson, 1993; Richardson, 1990) that making changes in either direction is no easy task.

Research Questions

Because the primary goal of this research project was to help teachers change their assessment practices, the primary set of questions addressed the effect of the staff development intervention on teachers' assessment programs—what did they try; what problems did they encounter; what advantages and disadvantages did they find in performance assessment; and, most importantly, what changes did they make?

Because we see assessment and instruction as inextricably linked, and because we were interested in the effects of changing assessment on instruction, we also examined teachers' beliefs and practice about instruction. A second set of questions asks about these beliefs and practices—what was the effect of the teachers' work on assessment on their instruction; what instructional changes did teachers make; what effect did teachers report the

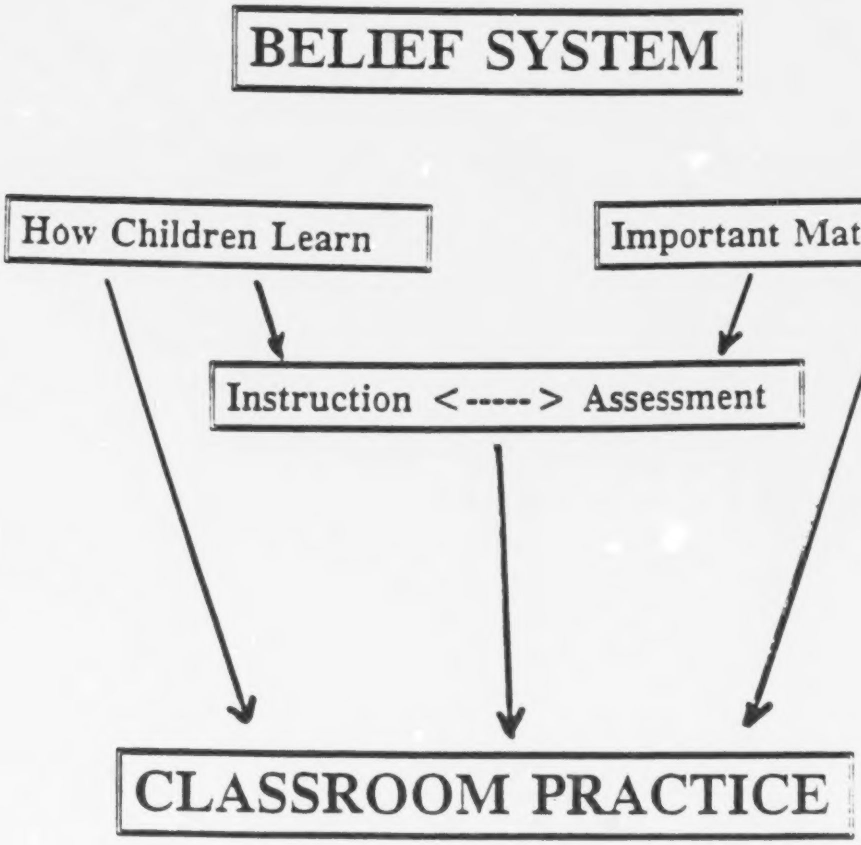


Figure 1. Knowledge and beliefs about how children learn and what mathematics is important to teach affect knowledge and beliefs about instruction and assessment. The three key areas are part of a teacher's belief system and will affect classroom practice.

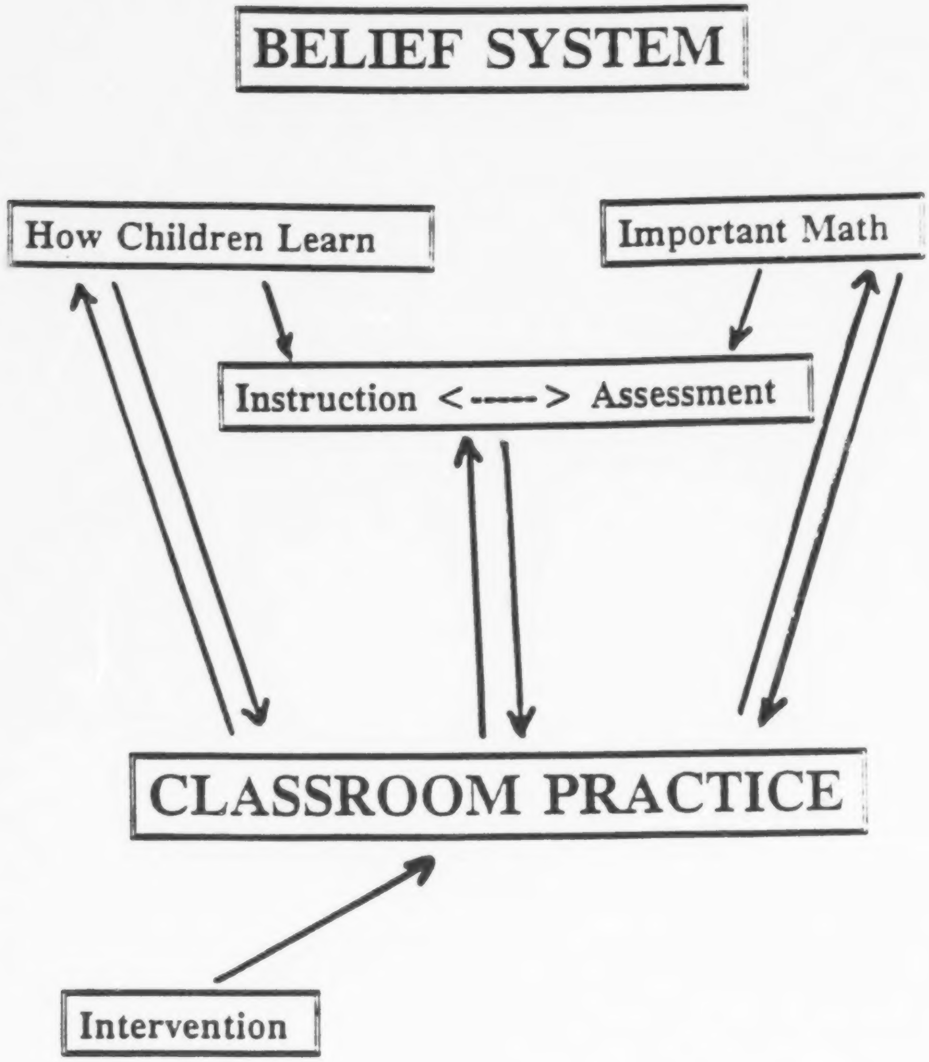


Figure 2. Applying an intervention that changes classroom practice can have an effect on a teacher's belief system.

changes had on children's learning; and how did teachers view the new instruction? And the questions that are very much a part of teachers' belief systems ask—what are teachers' beliefs and practice about how children learn; what is important to teach them in mathematics; and were there any changes in these beliefs or practices?

Method

The Project

This paper is based on data collected during the 1992-93 school year as part of the Alternative Assessments in Reading and Mathematics (AARM) project. The professional development aspect of the project was designed to help third-grade teachers select, develop, and improve classroom-based performance assessments in reading and mathematics that were compatible with their instructional goals. Our overarching research goals were to describe and explain the effects of these professional development activities on the instruction and assessment practices, and knowledge and beliefs of participating teachers, and on student outcomes. This paper describes the effects of staff development efforts in mathematics on several teachers with whom we worked. The team working with the teachers in mathematics throughout the year consisted of a mathematics educator, an expert in assessment, and a specialist in teacher change. The team had the assistance of several doctoral students and a visiting researcher.

Participants and Setting

We sought a school district that had a standardized testing program in place, a large range in student achievement, and considerable ethnic diversity. The district had to be willing to waive standardized tests for two years in the schools in which we worked.

The district selected is on the outskirts of Denver with a population that ranges from lower to middle socioeconomic status. The research team worked with 14 third-grade teachers in three schools (5 in each of two schools and 4 in the third). Each school submitted a letter of application signed by the principal, by the school's parent accountability committee, and by all third-grade teachers in that school.

While all 14 participating teachers were technically volunteers, some were less enthusiastic than others to engage in the project. Some of the original teachers who volunteered changed grade levels or schools and were replaced by other teachers who found themselves involved in a project for which they had *not* volunteered; others may have been "strongly encouraged" to volunteer. Our original assumptions were that all teachers were true volunteers and enthusiastic about the national reforms in reading and mathematics that their district also supported. We later found that these assumptions were incorrect.

Intervention

The intervention was a program of staff development, the primary vehicle for which was a series of weekly workshops between teachers and researchers; reading and mathematics were the focus in alternating weeks. The original intention of the workshops was to help teachers expand their classroom assessment repertoires, for example, by helping them learn to design and select activities, develop scoring rubrics, and make informal assessments "count." A second purpose for the workshops emerged early in the year. Many teachers requested materials for teaching in a way that their district now required and that would match the new assessments, so the scope of the workshops broadened to include more focus on instruction.

It also became clear early in the project that most teachers held fairly traditional views about what mathematics is important to teach, what instruction should look like, and how students should be assessed. Even teachers who were teaching or planning to teach in more activity-oriented, problem-based ways primarily used traditional tests of facts and skills for assessment. Because the instructional and assessment goals of the project matched those of the district (closely aligned with the NCTM *Standards*), we were at odds with the knowledge and belief systems of most of the teachers. Given that we were in the schools to help teachers with assessment, that the teachers had requested help with changing their instruction, and that we had not proposed a project to challenge beliefs, we took the position that teachers, like researchers, would learn from the evidence they accumulated from their classrooms. We worked on assessment (and instruction as teachers requested) in the context of current reforms in measurement and mathematics education, asking teachers to select and use instructional and performance tasks with their students and to bring feedback. We also worked with them on a plan for assessment for the term.

Our discussions in workshops were often about teaching with hands-on, problem-based materials and activities. The project provided tasks (see Appendix A for examples), many of which required problem solving, reasoning, and explaining, that could serve for both instruction and assessment. Because we had agreed to provide tasks that matched teachers' instructional goals and because those goals were primarily computational, most of what we provided the first term focused narrowly on place value, addition, and subtraction. The tasks were also short and structured so that teachers could see the connection between what they were teaching and the assessment task. One might say we were asking them to take small steps. We also selected tasks from sources that are easily available to teachers, so they would be able to make selections independently. We tried to help teachers think about their instructional goals, particularly what they want students to know and why; what it means to know math; how to tell if a student understands mathematics; and how to design and select problem-solving activities to elicit higher order thinking. Dialogue at workshops was about, among other things, selecting, extending, designing, and using activities and materials for instruction *and* assessment; making observations and how to keep track of them; analyzing students' work; and developing rubrics for scoring it. There was major emphasis on helping the teachers see the connection between assessment and instruction, that is, the "embeddedness" of assessment in instruction and curriculum.

The intervention or staff development included several full- or half-day in-service workshops attended by teachers from all three schools, the biweekly workshops within schools, project "assignments" that each teacher did with her class between workshops, demonstration lessons in two of the schools, and consultation on making observations in the third. Three interviews that were part of data collection (see below) are also part of the intervention because they gave teachers a chance to reflect formally on their beliefs and practices.

Sampling

A sample of six teachers, two from each of the three schools, was selected for in-depth study for this paper. The teachers were selected, after an initial analysis of the data, to represent a range of assessment and instructional practices and comfort with mathematics and mathematics teaching and were moderately to strongly engaged in the project. The method of selection, based on the initial analysis frame, ensured that the six cases are representative of 10 of the original 14 teachers. Of the remaining four teachers, one was marginally engaged in the project; the other three had more limited mathematical content knowledge.

Data Sources

The analyses for the present study were based on two sources of data collected from all three schools: semistructured interviews and biweekly workshops. All teachers participated in face-to-face interviews three times during the 1992-93 school year: fall, winter, and spring. The interviews were designed to assess teachers' knowledge, beliefs, and reported practices about mathematics instruction and assessment, as well as the relationship between assessment and instruction. A member of the research team conducted each interview; each interview took place at the participant's school during the day. The interviews were audiotaped and transcribed.

All 15 mathematics workshops from each school were read and coded (see analysis section below for description of the coding scheme). For the second round of analyses we then selected 6 workshops from each school,[3] 2 each from fall, winter, and spring, that addressed our project goals most explicitly and extensively. We decided, based on an initial analysis of the coded transcripts, that this sampling strategy would enable us more easily to search for trends without losing valuable information about patterns in the teachers' knowledge, beliefs, and practices.

Data Analysis

Our analyses began with all five authors reading the same two transcripts (one interview and one workshop) to develop a tentative coding scheme that would take into account issues of learning, instruction, and assessment in mathematics, as well as teachers' background and reactions to the project. This coding scheme went through two more iterations; that is, we coded different workshop and interview transcripts, discussed our codes, and modified the scheme. Our final coding scheme included categories listed in Table 1. Additionally, whenever a teacher talked explicitly about changes, we added a flag for change to the original code (see Appendix B for complete description of the coding scheme). If teachers mentioned change in an interview that did not fall under one of the original codes, for example, if a teacher talked about her growth in confidence, it was given a code for teacher insight or learning (Tlrn).

---

[3] For one school, 7 workshops were analyzed because each targeted teacher was absent from one or more workshops initially selected for in-depth analyses.

During the second stage of analysis, we developed "cases" of each of the 6 targeted teachers, that is, summaries of data for each teacher organized

Table 1

Coding Categories for Analysis of Interview and Workshop Transcripts

- Background Underlying Instruction and Assessment
  - Beliefs about students' learning
  - What it means to know mathematics
- Instruction
  - Teachers' goals for mathematics learning and instruction
  - Instructional tasks and activities
  - Organization and management of instruction
- Assessment
  - Roles and purposes of assessment
  - Content/substance of assessment tasks
  - Scoring of assessment tasks
  - How teachers keep track of what students know
  - How teachers assign grades in math
  - What teachers hoped to learn about assessment through this project
- Reactions
  - Dilemmas the teachers faced
  - Dilemmas the researchers faced
  - Advantages and limitations of performance assessments, including changes in student learning
  - Advantages and limitations of the project

according to several key areas. (At this point we focused on the three interviews and the sample of workshops, rather than the entire set.) These key areas were drawn from the original coding scheme by eliminating several less productive codes and expanding key ideas where our data revealed a rich picture about changes in beliefs, knowledge, and practices of these teachers. The three key areas were: (a) beliefs and practice about how children learn mathematics; (b) beliefs and practice about what school math is and what is important to learn and assess; and (c) beliefs and practices about instruction and assessment. These areas were augmented by data about variables that we considered important to this study: comfort with mathematics teaching, support for change, and engagement in the project. Because the area of beliefs and practices about instruction and assessment was central to our goals and included extensive data, it was divided into the following four subcategories: general instruction and assessment, problem solving, explanations, and additional assessment. Beliefs and practice varied from a "traditional" conception (e.g., children learn by being told; school math is about facts and computation; instruction is through the text; assessment is through tests of facts and computation) to a conception aligned with the NCTM *Standards* (1989, 1991, 1993) (e.g., children figure things out themselves; school math is about mathematical thinking, patterns, relationships, and explanations; instruction is through activities that require doing, thinking, reasoning, communicating, and generalizing; assessment is through multiple sources of data that give teachers evidence of student abilities to do, think, reason, communicate, and generalize). The variables of support, comfort with mathematics teaching, and engagement with the project varied along dimensions from limited or low to generous or high. (See Appendix B for more details.)

Our third and final stage of analysis entailed "looking across" these cases for themes that best describe the effect of the intervention on changes in this group of third-grade teachers' beliefs and practices about mathematics instruction and assessment. This final analysis addressed the research questions initially posed for this study.

Results

In this section we present themes that emerged within each of the three key areas from our analysis: beliefs and practice about (a) how children learn mathematics, (b) what school math is, and (c) instruction and assessment in mathematics. Although our primary interest is in the third area, we begin with the first two areas because of their influence on the design of instruction and assessment. We then discuss beliefs and practice about instruction and assessment and how teachers changed in these areas.

To protect their anonymity, teachers' names are not used, and the findings are presented in a way that prevents reconstructing individual cases.

Beliefs and Practice About How Children Learn

We found two major themes in examining teachers' beliefs and practice about how children learn. The first has to do with differences among children and the second with how learning should be structured in mathematics and the importance of children's comfort.

Differences among children. Most teachers believed that some children are more capable of doing mathematics than others. Teachers in this project believed that observed differences among children's mathematical capabilities are the result of either developmental differences at a particular time, or enduring differences in children's native abilities. One teacher compared learning mathematics to the way children learn to speak—at an early stage a child understands more than he or she can say, so the child has *received* concepts and information but is not ready to *transmit* evidence that she or he has them. Some teachers frequently reminded us that their students are only eight years old and may be at too early a developmental level for higher order thinking tasks, or at least that some third-grade students are not ready. Further, at least two teachers in the fall held the position that a few children in each class may *never* reach a developmental level that allows them to understand and should of necessity be taught by rote. For example, early in the year one teacher said:

> . . . a child like that, maybe we're better off just teaching him how to add and subtract on paper the traditional way, because that child may never until he's 30 understand what he's doing. See, I'm not sure that understanding has to come

> before doing it. I think many times doing it on pencil and paper, later then will help you understand it. See, I'm not sure that understanding has to come first. Because I think some children aren't capable of understanding.

She went on to say that most of the children *will* understand, and that she was talking about only a few. This teacher seemed to soften her position by winter, moving from the view that some children may lack *capacity* to the idea of developmental levels.

> there are children who just developmentally, aren't thinkers yet. And what we feed into them they can spit out, but they're not mature enough to really do a lot of real heavy thinking. . . . I think it can be, you know, developed, but some children are at different developmental stages and some kids just aren't ready for that. I have a couple of them in my classroom that just seem to, you know, if I show them how to do a problem, they can do it. But to really do some thinking about it, it's hard for them.

One teacher thought that some children had more logical ability than others and that would affect their capacity to do mathematics.

> . . . some children think more logically than others when it comes to everything and they are better in math and some children have no logical thinking at all and that is one reason why they just don't do well in math.

Teachers with either of these beliefs would be unlikely to present children with material, either for instruction or for assessment, that required higher order reasoning and problem solving—processes the *Standards* promote for all children. As the year progressed, some teachers were surprised at how much third graders could do and became more willing to increase their expectations. By spring, most had a view of the developmental continuum for third graders that included higher order thinking.

Teaching children in small steps and keeping them comfortable. A second theme involves how teachers believe children learn mathematics and also involves teachers' concerns for the comfort of their students. Most teachers believed that children learn mathematics by having mathematical concepts and procedures explained to them in small steps. Prior to this project, all but one of the six teachers had demonstrated their view of how children learn by telling, explaining, and showing, along with some questioning. They had, prior to this year, depended heavily on their textbooks to guide their instruction, holding the traditional view that children learn by being told and shown and then practicing exercises. Children's comfort was very important to the teachers, and this method of instruction appeared to be the path to comfort. For all but one teacher in the fall this meant presenting material in small bits and modeling carefully what the child was to do. For some this also meant that rote instruction of procedures was appropriate because understanding would follow the doing; that is, children learn "how" before they learn "why."

For several teachers, teaching students to do computations without understanding was also acceptable because doing procedures that others in the room can do would raise the student's self-esteem. Similarly, teachers were reluctant to give children tasks they might find frustrating. Yet, if children were used to being shown how to do everything, then any task requiring them to figure out what to do as well as to do it might cause discomfort. One teacher was ambivalent and was determined to give her students problems to solve and explain (even if, at the beginning of the year, "it made some cry"), *but also* to shape responses to problems to the point of eliminating most of the task's problem-solving character. For example, having selected a task that required students to find two-digit numbers that sum to 25, she gave the students the task with 3 sets of boxes set up as an addition/subtraction exercise.

> Because I really didn't think my kids were going to get two digits. I mean I didn't think they were going to understand the concept of two digits, and so I . . .

All of the teachers believed that experiential learning has some place in instruction, although at the beginning of the year only one teacher's primary mode of instruction was modeled after the position of the NCTM *Standards*. She seemed convinced that children could figure things out for themselves and that part of their work was to solve problems.

> I would see myself as most commonly, or probably the most often as the questioner posing questions, and then letting kids figure out how to work things to get an answer to that question.

Two others expressed a desire early on to move in this direction, although their later frustrations suggest they had not anticipated the full implications of this kind of instruction. Even at the end of the year, two teachers were concerned that children may be confused during hands-on activities and, unless carefully guided, may go through the motions without learning anything. One thought that some children are "dependent" workers and would be unwilling or unable to discover important concepts on their own. Even though she believed children learn from these experiences, she had doubts about using them.

> If they are dependent workers they need somebody to guide them through. They don't learn by the discovery method . . .

The implication for assessment is clear. If students must be told everything in order to learn it, then it is unfair to give them a novel or unfamiliar assessment task. If, however, teachers expect children to use their knowledge to solve unfamiliar problems, then an assessment task can present a problem for which no method of solution was taught. Teachers' reactions to the latter idea coincided with their beliefs about how children learn: from wanting to set problems that are challenging,

> I often look for problems that don't really have a solution. Sometimes I really like problems that have lots of solutions.

to wanting to narrow the tasks until the students knew exactly what they were to do. But even the teacher who wanted to challenge her students used assessment challenges that were within a reasonable expectation of what students could do. For example, when she was shown a missing-digit assessment task that involved regrouping, she modified it to one that did not.

Beliefs and Practice About What Is Important to Teach in School Mathematics

In the fall, we asked teachers what their overall instructional goals for mathematics were for the first quarter of the school year and then, over the year, asked them what they considered important for students to learn specifically about addition and multiplication. We also asked teachers in fall, winter, and spring what they mean when they say a student is "excellent" in math. Two themes emerged from these conversations about goals and questions about what it means to be excellent in math. The first was about computation, the second about problem solving and explanations.

Computation. All teachers talked about the importance of knowing and understanding facts, skills, and computation throughout the year. However, the emphasis was different for different teachers, and the views broadened during the year. In the fall computation was valued predominantly, but several of the teachers also talked about wanting children to be able to see patterns, estimate answers, and think about the reasonableness of answers. For one teacher computation was *not* a final goal, and even in the fall she said:

> . . . the computation that we do is really a means to an end. That [it] is not enough for you to be able to add three three-digit numbers. I mean, we want you to be able to do that, but that's not enough, they need to be able to apply it . . .

Another teacher whose major emphasis was on facts and computation in past years and in the fall was not as concerned about them in the spring. Facts and computation remained a primary focus for the other teachers, although their view of "understanding" a process broadened from expecting students to know that "3 X 4 means three groups of four" to expecting students to be able to explain, to show with models, and to apply the computation.

Problem solving and explanations. The second theme is that, as the year progressed, teachers gave more importance to strategies for problem solving and being able to explain how problems are solved and how procedures are done. Problem solving was mentioned at the beginning of the year as an important instructional goal for most teachers, but given the heavy use of the text, several teachers may have been talking about story problems. Teachers did not mention explanations as a goal in the fall, and one teacher may have expressed the concerns of several colleagues early in the year when she questioned the district's goal of explanation. In winter and spring teachers talked more about wanting students to be able to solve problems in real contexts. By spring teachers talked about knowing the difference between "problem solving" and "story problems," and "problem solving" had become an important goal, along with explanations.

Teachers' description of excellence in mathematics mirrored closely their instructional goals: a student who is excellent can do well all of the things a teacher listed as important to learn in mathematics. In the fall that meant he or she knows facts and can do computation accurately and quickly. Teachers also expected excellent students to catch on quickly, to be "good thinkers," and to be enthusiastic about mathematics. Teachers who valued problem solving in the fall included it among descriptors of an excellent student.

One teacher said in winter that there were two different ways a student can be excellent in math—either quick at computation *or* good at thinking and problem solving, but by spring she thought an excellent student would be both. By winter, teachers were also describing excellent students as those who could go beyond what had been taught, who sought challenging problems, and who might even make up their own problems. By winter, teachers also mentioned the evidence they expected to see from such a student—demonstrations of good understanding through explanations, writing, modeling, and problem solving. In the spring, all teachers talked about excellent students being good thinkers and skilled in solving problems and explaining their solutions; several teachers expected them to be able to produce more than one solution to a problem, and at least two teachers talked about students' ability to apply what they know to real world problems. There is evidence from their conversations in workshops that every teacher would have this latter expectation, although she might not have mentioned it specifically in the interview. In other words, just as the teachers' ideas about what is important in mathematics developed over the year, so did their view of what it means to know or be excellent in mathematics. Not only did their comments broaden to include more higher order thinking, problem solving, and explaining, but they showed a keener awareness of the evidence they can collect as proof of these processes.

The implications for assessment and instruction of a teacher's ideas of what is important to include in a school mathematics program and what comprises excellence in mathematics are clear. When the emphasis is on computation (as it was for most of our teachers in the fall), then classroom tasks reflect that. When teachers value mathematical thinking and problem solving (a shift we saw in most teachers to some extent by spring), both instruction and assessment will include activities that require students to think and solve problems.

Instruction

Even though the primary focus of this research project was on assessment, we became interested in instruction for three reasons: (a) We believe instruction and assessment progress in tandem; (b) advocates of performance assessment claim beneficial effects on instruction; and (c) the teachers requested assistance with their instruction.

Teachers were asked specifically about their instruction in interviews in the fall, winter, and spring. They also talked about their instruction frequently in the workshops and shared with the research team classroom activities and methods they were using. Three themes emerged: (a) Teachers changed their instructional practice; (b) teachers perceived that students had learned more; and (c) making instructional changes was difficult.

Shift in instructional practice. There was a shift during the year toward using manipulatives, hands-on small-group activities, problem solving, and explanations; and, for the four teachers who used a text in the fall, a corresponding shift away from it. One of the teachers had been teaching in this way before the project started, so that her shift was not so striking, but by spring she was doing more problem solving and requiring explanations that she had not required before. For the teacher who called the text her "bible" the change was dramatic. The shift away from the text surprised two other teachers who had been convinced that their text was excellent. They initially saw no reason to leave it and supported it vigorously to the research team. But when they compared it to the district's new goals for mathematics, they saw the inadequacies of the book, both in coverage of certain topics, for example, probability, and in the book's approach to teaching. They continued to use the book as a source of exercises but shifted to more activity-based instruction.

> [We] found holes in the text book so we used a variety of resources in order to build a unit around probability and statistics. And we spent a whole, the whole grade level, . . . created centers for probability and statistics, and then we exchanged those and we did it with whole group and the kids were, had a variety of materials, spinners, colored, colored tiles . . . dice and we found that in our book there was only one page on probability and statistics. And that is an important strand.

By spring all teachers reported having students solve more problems, write more explanations, and engage in more hands-on activities and suggested that the set of resources our project had supplied facilitated this change.

An interesting, unplanned curricular development became an influential addition to our intervention. Teachers at all three schools adopted the Marilyn Burns multiplication replacement unit, *Math by All Means: Multiplication, Grade 3* (1991). For one school team the project year was the second year of using the Marilyn Burns unit, but it was a first experience for the other two school teams. In one of those schools, the unit was used by the math specialist at the school; the classroom teachers did some follow-up but only one teacher at the school, one of the two in our sample, was significantly involved. Although all teachers mentioned some use of manipulatives in the fall, for several these were limited or largely nonsubstantive; for example, a child could roll a pair of dice twice to get the two numbers he should add together. The Burns unit gives a teacher complete instructions for a hands-on, manipulatives approach to teaching multiplication that includes solving problems and explaining answers and solutions.

This unit may have had considerable effect on the teachers at the first two schools and the one teacher at the third. Teachers had a model of exemplary nondidactic teaching, and they saw how it engaged students. It showed them a way to use manipulatives that was not routinized, although we had discussions with some of the teachers about whether or not students could go through the activities in a rote and mindless way. This unit used manipulatives as models for computational processes, and some of the models were new to most teachers, for instance, rectangular arrays of tiles to represent the product of two numbers. The multiplication unit seemed to make most of our six teachers more comfortable with substantive, hands-on learning; some, of course, already were.

Beyond the multiplication unit, the areas in which teachers felt most comfortable exchanging the text for hands-on activities seemed to be those that were noncomputational and had not been stressed in their programs in the past. For example, teachers at one school developed their own unit on probability, organized around menus of activities; and all three schools used hands-on activities to teach geometry.

We saw some exciting changes in a teacher who had vigorously resisted many of the project ideas. She talked about changing her instruction because of the assessments, and how using the Marilyn Burns multiplication unit along with the activities provided by the project had made her see

> how you change your instruction so that you're making children think more, more engaged, relating it to their everyday life.

She talked of the project being a "catalyst for change," and said that even though the anxiety it produced was not always comfortable, anxiety is sometimes necessary in order to get change.

A teacher who had taught very traditionally in the fall got lots of positive feedback from seeing how much her students now enjoy math. She said:

> T: I like math better myself.
>
> I: Why do you like it better?
>
> T: I just like the way I'm teaching it. The kids are enthused about it. I make sure I have math everyday. Last year, I can't say that.
>
> . . .
>
> Yeah, last year I'd skip a week or two. But the kids do ask for math; they like math.
>
> . . .
>
> I'm doing a better job this year.

Student learning. Teachers reported that they thought their students were learning more and had better understanding. By the end of the year students could solve problems and give explanations at a level that surprised many of the teachers. Teachers were stressing flexibility in solving problems, and students were responding with multiple approaches to their solutions.

> T1: Well, I just think they understand it more, it is not just rote memorization—that they really know what it means when you say 20 times 80 even if they don't know the answer . . . There is a much deeper understanding.

> T2: But I think we have given a lot more challenges this year to our group that we would normally not have given a normal third grader. Don't you think?
>
> I could say that she's been exposed to a lot more problem solving than she would have been in my classroom last year.
>
> T3: Also something I'm really encouraging with my kids is to be flexible, that there isn't one way. Today we solved a problem and we got six different explanations of how you could have possibly solved it. In my mind, math has been, in the past, right or wrong, and I'm really trying to encourage them to think flexibly, to be flexible in their thinking that, well if it didn't work this way I could try this, or if it worked this way could it work another way? Could I look at it from a different avenue?

Difficulties with new instruction. The third and not surprising theme is that some teachers had difficulties with two aspects of this kind of instruction. One aspect involved content. Teachers were concerned, for example, with the Marilyn Burns unit, that students would not come away with knowledge of facts and appropriate skills. While they agreed that students had a better understanding of multiplication and its application, they questioned whether it taught the facts adequately and whether students were learning anything from all the activities.

> . . . how to use—to do menus independently and a lot of them were going through the motions of it but they weren't catching multiplication.
>
> . . .
>
> Yeah, other people liked it. But, I had to make a professional judgement. Now I will do Marilyn Burns again but at the same time I will be working—I will incorporate the multiplication tables at the same time. When we were done with Marilyn Burns I think maybe they did have an understanding of multiplication, what we were looking for . . . [but] they can't do any of their tables, then I had to take four weeks out of my math curriculum to work on the tables.
>
> (Oh, so they didn't know any of their tables?)
>
> They didn't know any tables, but I think they had a basis for—that's why we will go back to it. I do think they had some multiplication understanding of the real world,

> like they looked at things in multiplication. They looked at egg cartons and they saw that things came in sixes, where before I think I just taught the multiplication tables and they never related it to the real world.

The other aspect involved the organization of instruction alternative to the text. As already discussed, two teachers thought their text excellent and saw no reason to change, particularly when it was all organized; leaving the text requires planning, collecting, and organizing new materials. It is unreasonable to expect teachers to choose to add burdens of curriculum development to those of teaching their classes. Even teachers who had been given materials for hands-on instruction in courses they had taken needed time to organize them.

> I have taken all of the math manipulative courses in the district so I got that [a set of activities] from [a district math specialist]. So I was very familiar with them. But I never—it just takes some time to fit it all in, like when to use it and how much do you run off, and you really need that, and then being able to make a critical viewpoint of how much we need and the variety of levels, being able to read that.

Although most teachers welcomed the resources provided by the project and found them useful, these resources themselves increased the amount of material with which teachers had to cope.

All of the teachers found the additional work in the project burdensome in the fall, and by Thanksgiving, they were feeling overwhelmed. The project director negotiated arrangements to ease the burden, for instance, a half day each month of released time and only one weekly assignment instead of two (one each for math and reading). For many of the teachers these arrangements seemed to remedy the problem. Of course it was also the case that they were becoming more comfortable with the new assessments. A couple of teachers remained frustrated, particularly if they were trying many new practices. For example, one teacher had enthusiastically embraced the kind of instruction and assessment we, her district, and NCTM were advocating and set out to revamp totally her mathematics program. By February, she appeared to be overwhelmed with the magnitude of the changes she expected of herself and was having second thoughts and returning to worksheets.

> I am giving more worksheets at this point in time because I found that I couldn't just do problem solving and there needed to be a point in which I went through the same old steps I had done before.

> I feel that it needs to be a little more structured than I had it in the fall. Because we're doing the new significant learnings I kind of jumped into . . . this manipulative and problem solving and no worksheets. But I find there has to be a balance. You can't throw out all the stuff we used to do. Even for your own sanity you have to have some of those things like that [worksheets] while you're getting used to the new program.

Spring found her proceeding with caution, doing more problem solving, but continuing to present material in small steps for her students.

This teacher was not alone in talking about wanting to keep a balance among facts, computation, and problem solving. The actions of all the teachers and their comments about what they valued in school mathematics suggest this was something they all thought about. The balance was, of course, different for each teacher. The most vocal seemed to be telling us we were trying to pull them toward problem solving to an uncomfortable degree; they were also the teachers whose programs had had the least emphasis on hands-on activities and problem solving.

> I personally, I still feel like I need a balance of both. I don't want to do all problem solving everyday, this kind of problem solving. And I don't want them to do all pages out of their books everyday. But I do think for them to survive, I think they need a balance, and I want them to be able to do some thinking skills, but I also, if they go to fourth grade next year and the teacher says you need to do page 36, 1 through 25, I don't want them to look at each other and not have a clue on what they would do with something like that . . . not know how to put a heading on their paper or write their numbers so that they can be read by other people. I think they need those things from that kind of practice no matter how well they know their facts from playing cards. I just think there needs to be both. I think they need to be able to write problems on paper and have somebody else be able to read them.

Assessment

A set of themes corresponding to instruction emerged for assessment: (a) By the end of the year, teachers were using more authentic evidence to assess what students know; (b) in spring, teachers reported knowing more about what their students know; and (c) (again, no surprise) teachers encountered many difficulties with performance assessment.

Shift in assessment practice. The first theme is the central goal of this project—to help teachers select and/or design performance assessments that expand the variety and quality of ways in which they assess their students. Because established policy at all three schools required timed tests of facts, all teachers used such tests during the year, but some more frequently than others. One teacher's fall program included daily one-minute tests of facts. All teachers also graded children's work on daily computation during the fall, either from the text or from a set of five problems written on the board. At least one teacher in the fall graded students' daily work for neatness and format as well as for accuracy. The teachers described earlier, who valued their text in the fall, also used its pre and postchapter tests (parallel forms of the same test), although they used them differently. One gave the pretest at the beginning of the chapter's work and the posttest at the end to show both the students and the parents how much the children had learned. The other gave the pretest a few days before the posttest at the end of the work on that chapter, more as an instructional and diagnostic device to help students do well on the posttest. Note that she is one of the teachers who is concerned about the comfort level of her students, and this test preparation probably provided a level of comfort as well as training for the "real" test. But however and whenever these paper-and-pencil assessments were used in the fall, the major focus was on recalling facts and doing computation. The pattern began to change by winter.

The early work in the math workshops was about assessing important mathematical skills, broadly defined, as in the NCTM *Standards*. The research team encouraged teachers to assess more broadly—that, in addition to competence with paper-and-pencil computation, it is important and useful to develop and assess children's ability to model numbers and procedures, make estimates of them, explain them, and solve problems about them. By winter all the teachers were trying to be more systematic in their observations of these abilities and were using problem-oriented computational tasks to assess them. They were requiring children to give explanations, both orally and in writing, of how they were performing procedures. For example, teachers gave students problems with missing digits to solve and to explain their solutions; they also gave them "buggy" problems to do and explain.

(See Appendix A for examples of tasks teachers were given to try; see Appendix C for examples of their assessments.)

The assessment of students' work on these problems in the winter was still at an informal level; that is, they were not scored and recorded in the grade book, merely noted for the information they provided about students. In addition to these more alternative tasks, most teachers continued to use some form of computational tests, either daily pages from the text, examples on the board, or chapter tests, and scores from these *were* recorded in the grade book. It was almost as if the alternative kinds of assessments were interesting activities for children but did not have the same weight for assessment as a computational test. This began to change in the spring.

One focus of the winter and spring math workshops was the scoring of students' explanations, both for explaining procedures and for explaining their methods of solving problems. Teachers developed a variety of general, and very brief, rubrics and applied them to students' work. By spring, all teachers were using students' problem solving and explanations for assessment, although two expressed concern that a child's problems with writing might mask his or her mathematical performance. Even so, all teachers adopted assessments that require written explanations, and they all noted that it was one of the major changes they had made this year. Two teachers tried to deal with the problem of poor communication skills by giving two scores—one for the answer and strategy used and the other for the explanation of the solution.

> And I found that for some, for many kids there are a lot of times [there's] a big discrepancy in whether they had a good strategy and whether they could really explain all of that strategy. And so I have now divided up my marking, a viable strategy and an explanation. Because I thought some kids need credit for their thinking even though they didn't write it out in words, but it's obvious to see the

> thinking that . . . Because like with [student] now, I mean there was nothing written, but actually after he told me the words I made sense of his picture.

Two teachers talked about giving a daily problem for "experience" but scoring only one each week. One of these teachers required students to write explanations only for the problem to be scored, while the other insisted that students write explanations daily. At least three teachers asked children to score their own and classmates' explanations for the instructional value it provided. As children worked on scoring explanations and saw many examples, they were more likely to internalize the criteria.

Even in the fall, all teachers talked about observing and questioning children, for instance, "Show me five groups of three." They all knew that these observations and exchanges were sources of valuable information about their students' understanding, but seemed not to consider them part of their program of assessment. Only one teacher kept systematic notes; and only one other expressed a desire to systematize her intuitions about what students know, and she placed the highest priority on learning how to make systematic observations. She also felt that she knew what each child knew but wanted to verify her "gut feelings." In fall she said:

> I'd like to be able to have more assessment that will give me some data to go with the gut feeling that I have. So that I could prove an understanding or a lack of understanding.

She also wanted checklists for proof of what children know and to help her plan instruction. In winter, her response to an interviewer's question (Why do you want checklists?) was:

> I think for proof. I think that if someone questioned me, you know if a parent said, well why, why this grade . . . either high or low, that I could say . . . well you know on this date when we were doing this, this is what I saw him do. . . . I think that it would be helpful to me too, to be able to after a lesson, just at a glance, look and see where kids are falling so that, you know, tomorrow I can maybe go to those kids first that are showing a weakness. . . . and one of the things that I find hard in math planning, is planning for a week at a time. Because what we do tomorrow depends on what happened today.

Two teachers were actively opposed to taking notes on these observations. They felt able to keep track mentally of where each student was and saw systematic recording of notes as cumbersome and burdensome.

In order to develop the assessment potential of observations, we made them another focus of our winter and spring workshops, primarily working on developing schemes for keeping systematic notes about students. Teachers developed checklists, used class lists with space for writing, drew grids with children's names in boxes, used spaces in their grade books for checks and other symbols, and even tried to use a copy of the assessment framework for each child to record how they were doing. All expressed frustration and doubts about these attempts. Sometimes a teacher's teaching style affected her ability to keep notes. Those who used direct teaching to the whole class had problems making individual observations. Those who had activity-based classes had difficulty getting around to each child and felt they wanted to give instruction every time they encountered a child with a problem. Some teachers who saw little value in systematic observation notes at the beginning of the year never became convinced of their value but felt they watched children carefully enough each day to know exactly who knew what and what difficulties they were having.

By spring, most of the teachers were trying to use systematic observations, some more successfully than others, but no teacher finished the year with a system for keeping anecdotal records that she felt worked well. The two teachers who tried to take systematic notes while observing children were overwhelmed by the amount of data they had for each child. They realized that anecdotal notes they had made could not be reduced to numbers recorded in a grade book. They thought perhaps that more selective assessment might be a solution for keeping the amount of data manageable. Two teachers seemed equivocal but convinced that they could keep the relevant information mentally.

Also by spring, the two teachers who had been using chapter tests were no longer using them routinely. One used no chapter test all spring, and the other said she used them only after critiquing them and judging them to be relevant.

(But you also said you used the chapter test or some part of it.)

> Yeah, but now I am looking at it more critically. Before it just used to be part of the routine. I look them over and if I feel that they are relevant I use them. If I feel that they are not relevant I just move right on.

These teachers and one other seemed to prefer a balance between traditional and alternative forms of assessment, partially because the alternative assessments the teachers developed had some ambiguities in the directions.

> T: But I still think it needs to be a combination.
>
> R: What combination?
>
> T: Normal assessment and alternative assessments. I would never recommend to a classroom teacher to go with all alternative assessments.
>
> R: That's fine, and what are normal assessments for you, paper-and-pencil, computation?
>
> T: All these were paper-and-pencil.
>
> R: But see I look at, yeah so that's why I'm asking, what's normal? Is normal a chapter test, is normal computation?
>
> T: Like a standardized, a more standardized test because I think as we discover when you make tests there're always glitches in it. You know we've discovered that haven't we?

Also, teachers seemed more comfortable using new forms of assessment in the new instructional units they were trying, such as probability and multiplication. For the latter they were willing to select items from the Marilyn Burns unit and from tasks supplied by the research team; teachers at one school designed an assessment that was similar to the tasks they had developed for a unit on probability. Teachers' willingness to use performance assessments with unfamiliar topics occurred later in the year when they were becoming familiar with this kind of assessment, so it may be that as their comfort level rises, teachers would elect to use alternative assessments even with standard topics.

What is clear about the spring is that teachers were using many more forms of assessment than they had used in the fall, and that the nature of most these assessments had improved. They were focused more on children's thinking and on their performance on higher order skills. Teachers were observing children more carefully, and most were attempting to keep records of what they saw and heard. Most were willing to design their own assessments (with the help of their school team) even if only selecting from a set of tasks supplied by the research team. This was a change from fall when several teachers had been resistant to developing assessments, saying, understandably from their perspective, they did not care to "reinvent the wheel." One teacher was exceptional in her interest in and willingness to design many of her own assessments—some were extensions of those she was shown, and others were original. She also adapted an attitude measure from one she had for reading.

Teachers' knowledge of students. The second theme related to assessment is that teachers knew more about their students from performance assessments. Most teachers claimed performance assessments gave them new and deeper insights into children's thinking and understanding. They saw them providing much more information than whether a student can or cannot do something or whether a student "has it" or not.

> T1: . . . Whereas before we were doing all of it but didn't, we didn't have them, the samples of work, we didn't have the collections and I think . . . even our kids have a better understanding of what we expect and what we're looking for that kids previously didn't.

> T2: Well, I just don't think I ever really thought about math in terms of writing. It was more a numerical process, and I think being able to see how the kids explain through writing told me a lot about what they know and about their thinking process . . . kind of goes beyond the work sheet . . . be able to explain—not just answer but be able to explain it. It tells me a lot about them as thinkers . . . Just, I think, getting the picture of a math student as a whole and not just one part of math, can they add on paper and subtract and multiply—it just goes much further than that.

R: Have you learned things about students' knowledge of mathematics that you otherwise might not have learned as a result of these assessment strategies?

T3: Yes, mainly that they can understand and explain to me what they are doing. Otherwise I would I just assume that they knew.

T4: Advantages? Um, I think through the assessments that we've been working with, children can . . . can . . . I mean you can, you can see if they're really understanding the process . . . much more so than just, you know, rote learning and doing what you're supposed to do.

. . .

I think you see how they are thinking . . . and how they problem solve better.

Difficulties with performance assessment. The third theme, that teachers had many difficulties with performance assessment, came as no surprise. The problems teachers faced were understandable and were proportional to the amount of change they attempted. Initially, difficulties had to do with lack of knowledge about what a performance task was, how to use it, and how to score it; and with observation, how to acquire and keep track of information about individual students and teach 25 others at the same time. We discussed above some problems teachers had with systematic observations and with scoring explanations, but they also had problems of a more general nature. For example, there were some initial misunderstandings at one school about teachers' perceptions of "teaching to the test," something they wanted to avoid. The teachers' interpretation was that their assessment tasks had to be very different from the performance tasks they had selected for instruction, and so, after using a wonderful set of instructional activities to teach place value, they chose a set of traditional worksheets for assessment. In addition to their misunderstanding, they believed then that paper-and-pencil computations were the definitive assessment for showing students' understanding of regrouping.

Teachers found it overwhelming to attempt changing their assessment program at the same time that they were changing their instruction in two major curricular areas (mathematics and reading).

> So, I feel like I could do such a better job and I said this thing before, if I was doing all reading this semester and all math next semester. I just think it would make it so much more manageable and I could focus so much more. I find myself going through the folder and I'm looking for what I need to have ready for you on Tuesdays and what I need to have ready for Freddy [the reading expert]. You know, I just, it's been a real management nightmare.

In the fall, many of the teachers saw the new assessments we asked them to try, and the new instructional activities they had requested, as add-ons to their regular instruction and assessment programs. Since they were trying to teach and assess everything as they had been doing, it was difficult to find the time to add the new instruction and assessments. And the assessments themselves took longer: Children take longer to solve a problem and write an explanation than to add some numbers. Scoring was also more difficult and more time-consuming: Rather than merely marking an answer correct or incorrect, each solution and explanation had to be read carefully enough to be scored. Another problem for one teacher was that scoring solutions to problems and explanations was too subjective and lacked the reliability of a standardized or chapter test from the text. Another felt performance tasks did not focus sufficiently on whether students know the facts and have computational skills.

The issue of children's comfort came up as a problem in these assessments, a concern we discussed earlier with respect to instruction. When children are given a problem as an assessment task, and they are not sure of how to solve it, they may be uncomfortable; they may ask many questions; they may whine; they may become unruly; some may cry, particularly if they have never felt the frustration of not being sure how to proceed. By training and selection, a teacher's response is often to want to tell children how to do things and to make them comfortable—just the opposite of what we were asking of teachers. By spring, most of our six teachers had adapted problems to their classes so that the level of difficulty was manageable, and they were rewarded with students who were enjoying the challenges. The early conversations about not giving an assessment task to a student unless you had shown the student how to do it were no longer heard in the spring.

Several teachers mentioned concerns about what parents might say if they did not send home tests of computation and if they used performance assessments instead. Despite the findings of another part of this study

(Shepard & Bliem, 1993) that parents were overwhelmingly in favor of performance assessments, teachers feared that that would not be the case. Another teacher expressed surprise when parents were receptive to her including students' performance in solving problems as part of their grade. The resistance of their colleagues in higher grades to their working on mathematics other than facts and computation was also a problem for several of the teachers. Each school had a policy of requiring a certain score on timed tests of facts by the end of each grade, and this requirement seemed to hang heavily as a responsibility on most of the teachers. It is clear that the support of other teachers in the school and parents was important to have, and lack of it, real or perceived, was distressing to teachers.

> It's real frustrating because I know what the thinking is and I know what, pretty much what we're supposed to be doing. But then I was talking to a fifth-grade teacher the day before yesterday and she was saying how the kids don't know their facts and they can't do their computation skills. It's like we're being geared to do problem solving with the kids and all that, and then teachers in upper grades are upset because they're coming into them and not having the computational skills that they think they should have. One teacher does math timed tests and we hear, "No we shouldn't be doing math timed tests, that's not a valid way for kids to learn their facts." It's like being pulled in two different directions. And we can teach the problem solving and, at least we're trying to be able to do that. Not all people believe that that's the way—what we should be doing—and then we send our kids up to them, and it's like, (Could this child do their timed tests when they were in third grade?" Do you know what I mean? Don't you guys feel like that, like you're being pulled in two different directions and then parents come in and say, "I don't understand why my child doesn't bring home 25 addition problems every night to work on, what good is this going to have them do to count the legs on this animal."

It appeared that strong grade-level support was important and helpful to teachers, although even with such support, a teacher could still find the suggested changes too difficult to make. On the other hand, lack of team support did not appear to disturb another of our teachers, as she made significant changes in her instruction and assessment programs.

The difficulties teachers had with performance assessment were similar to those of making any change—not understanding how to do it, not having the time to take it on, thinking they had to add it to what they already used, being overwhelmed by what they were trying to do, doubting whether the change was sound, seeing that the change made their students uncomfortable, and feeling they lacked the support of other teachers and parents.

In summary, the effects of the first year of our project on teachers' practice of instruction and assessment were numerous. Teachers were using more hands-on activities, problem solving, and explanations for both instruction and assessment by spring. They were also trying to use more systematic observations for assessment. All teachers agreed that their students had learned more that year and that they knew more about what their students knew. Every teacher struggled with the revised instruction and new assessments, even those who endorsed them most enthusiastically. Many of the teachers used the word "overwhelmed" in referring to how they felt during the year, but they responded to feedback from their own classes about performance assessment and activity- and problem-based instruction. The feedback they got was generally positive; for example, their students seemed to have more conceptual understanding, could solve problems better, and could explain their solutions. Teachers' response, for the most part, was to attempt further change in their assessment and instruction practices and to become more convinced of the benefits of such changes.

Discussion and Conclusions

This paper reviews a year of work with third-grade teachers during which performance assessments were introduced in order to improve both instruction and assessment in mathematics. The major finding of the study is that participating teachers adopted many changes in their instructional practices (with respect to content and pedagogy) and their assessment practices (with respect to methods and purposes). Moreover, changes in assessment and instruction were, for many, mutually reinforcing. By year's end, many were using more hands-on and problem-based activities more closely aligned with the NCTM *Standards*, as intended by the project, to replace and supplement more traditional practices of text-based work, and they had extended the range of mathematical challenges they thought feasible to attempt with third graders. They used more varied means of assessment, for example, performance tasks and observations, that either replaced or supplemented computational and chapter tests. One teacher whose instructional practices already reflected NCTM *Standards* made even more progress in that direction, and she was able to adopt more authentic assessment practices.

In short, the introduction of performance assessment provided teachers with richer instructional goals than mere computation and raised their expectations of what their students can accomplish in mathematics and what they could learn about their students. There is a certain irony in teachers' concern with their students' comfort and their awareness that solving problems made students less comfortable than learning and performing computational algorithms. One of the goals of the *Standards* is to empower all students mathematically and to make them comfortable with mathematical thinking and problem solving. It appears that to accomplish this long-term goal, students may encounter some initial discomfort.

We list in the results section the many problems teachers reported as they realized the magnitude of the task of revising both reading and mathematics assessment. Then, as most teachers realized they also had to revise their instruction to prepare students for the new assessment tasks, they felt overwhelmed.

It is likely that most teachers also felt uncomfortable with some of the changes, and with being at odds with recommendations of the *Standards*. The teachers, as we would expect, adapted differently to the challenge of change. We can use a Piagetian model of assimilation and accommodation to describe teachers' reactions. Those changes in practice that fit a teachers' system of knowledge and beliefs were assimilated into that system. So a teacher whose belief system corresponded to the district goals was able to assimilate new practices without discomfort, for instance, making anecdotal notes about students. She was comfortable with the task and had to deal only with the amount of work it implied (still a chore, but not an onerous one).

Other teachers also assimilated practices into their belief systems, even when those practices appeared to be discrepant with their systems. They simply adapted the practice to fit their system; for instance, a teacher who believed children learn by being told would show children how to use base ten blocks in a directive manner. These teachers also felt little discomfort, but had the work (again, no small amount) of selecting and adapting the practices that could fit. For some of these teachers the discomfort came with having tried to make too many changes.

The teacher quoted above, who said, "I know pretty much *what* we're supposed to be doing . . ." had not incorporated *what* into her knowledge and belief system. It was still something being imposed from the outside, and so when she met resistance from other teachers and had her own doubts as well, she pulled back from that kind of teaching. She could try some things in a superficial way, but if they had no comfortable place in her system, she was not ready to modify her system.

Practices that made teachers uncomfortable were sometimes rejected, for example, letting students cope with a problem they had no idea how to solve. But if there were reasons why the practice continued to be attractive, the teacher was drawn in two directions (the disequilibrium Piaget talks about), and she began to change her system of knowledge and belief (Piaget's accommodation). We saw an example of accommodation in the teacher who talks about the project being a catalyst for change.

While we did not try to change beliefs directly, we know we affected beliefs through changes in practice. There is no doubt that changes in beliefs alter practice, but it is also the case that shifts in practice may lead to shifts in belief which can, in turn, further affect practice. In this study the changes that teachers made were likely at first to be changes in practice. We saw teachers whose students gained greater understanding of multiplication from many hands-on activities change their belief about how to teach multiplication. As teachers got positive feedback from students about changes they had made in instruction and assessment, they were encouraged to attempt further changes. In other words, changes in beliefs and changes in practices appear to be mutually reinforcing. While this cycle appeared to lead to, for some, a fundamental change in instructional and assessment *practice*, it is not yet clear whether it also changed their *beliefs* about instruction and assessment.

We report many changes that teachers made in this project. What we cannot know is how durable or ephemeral those changes are. We know that some teachers made some changes superficially, adapting them to "fit," but other changes were made at more fundamental belief levels, and those will likely endure. Our work at two of the schools this year gives us confidence that, with continuing support, teachers are making even more changes. But, the question of the stability or persistence of the changes cannot be answered in real time.

What is abundantly clear is that the change that occurred did so not from anything we *told* teachers to do, but from their experiences with the ways performance assessments improved their classrooms. Just as we hope teachers will permit students to construct their own meaning from mathematical experiences, we must permit teachers to construct their own meaning for performance assessment.

It is important to ask if our intervention is a model for others. Not only was that not our intention; it is most unlikely that the number of personnel (four university faculty, seven graduate students, and one visiting researcher) devoted to work with 14 teachers could be replicated in a school district. Like the teachers, we were also "messing about" with how to help teachers construct new views of assessment, and through that, of instruction and learning. There are things we would do differently and some other things we hope to try next year (the third year with these teachers), for example, administering some larger performance tasks at the end of this year, perhaps from the Maryland assessment, and then discussing student responses with teachers the following fall.

We learned some things about what and what not to do, and perhaps staff developers can benefit from our struggles and experiences. We know that teachers need a lot of support (from experts, administrators, peers, and parents) for changes they are expected to make, and they need to have some reason for wanting to make them. They need permission to go slowly and perhaps make what might seem to be quite small changes, and to be able to make them over a period of time measured in years, not months. Teachers need many chances to try things out with children (to mess about) and help in discussing and interpreting their classroom experiences. They need a lot of encouragement for all the extra time and hard work it takes to make changes. Staff developers must expect to see stops and starts, and even occasional backward motion. They need to remember that all teachers are *not* at the same starting point; that the same intervention will not work for all teachers; and that each teacher will adopt different changes that match her or his existing beliefs and practices. Staff developers need to know that change in instruction and assessment is not an all-or-nothing proposition—that teachers have it or they don't (or even that everyone agrees on what "it" is)—and that teachers can comfortably hold inconsistent views and engage in inconsistent practices for a very long time. Finally, they can also expect to see some teachers who don't want to play and will want to sit this one out, believing about performance assessment that "this too shall pass."

In conclusion, our results are not a clean sweep. They show it is not a matter of "show the assessment tasks, and teachers will use them," nor is it a matter of "have teachers use performance assessment, and they will change their instruction." Nor are we making an argument for high-stakes enforcement of externally mandated performance assessment. It's not about forcing. It's about a lot of slow, often painful, hard work for both teachers and staff developers. It's about the delight when the teacher who argues most vigorously about the changes says,

> I've changed my instruction. . . . I mean I have to; I mean if I'm going to assess kids differently, I have to teach differently.

References

Battista, M. T. (1994). Teacher beliefs and the reform movement in mathematics education. *Phi Delta Kappan, 75*, 462-470.

Borko, H., & Putnam, R. (in press). Learning to teach. In R. C. Calfee & D. C. Berliner (Eds.), *Handbook of educational psychology.*

Burns, M. (1991). *Math by all means.* White Plains, NY: Math Solutions Publications and Cuisenaire.

Cobb, P., Wood, T. Yackel, E., & McNeal, B. (1992). Characteristics of classroom mathematics traditions: An interactional analysis. *American Journal of Educational Research, 29*, 573-604.

Flexer, R. J. (1991, April). *Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Gipps, C. (Ed.). (1992). *Developing assessment for the national curriculum.* London: Kogan Page.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Mathematical Sciences Education Board. (1989). *Everybody counts.* Washington, DC: National Academy Press.

National Council of Teachers of Mathematics. (1980). *An agenda for action.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1993). *Assessment standards for school mathematics—Working draft.* Reston, VA: Author.

Nelson, B.S. (1993, April). *Implications of current research on teacher change in mathematics for the professional development of mathematics teachers.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Seattle.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. *Educational Researcher, 19*(7), 10-18.

Romberg, T., Zarinnia, E., & Williams, S. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions.* Madison, WI: National Center for Research in Mathematical Science Education.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72*, 232-238.

Shepard, L. A., & Bliem, C. L. (1993, April). *Parent opinions about standardized tests, teachers' information and performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Shepard, L. A., & Cutts-Dougherty, K. (1991, April). *Effects of high stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Smith, M. L. (1991). Put to the test: the effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 71*, 703-713.

Appendix A

Math Tasks Provided by Project

R.J. Flexer
CRESST Project
C.U. - Boulder

Place Value, Borrowing and Carrying

1. Put 4 different one-digit numbers in the brackets to make
   - the largest possible answer
   - the smallest possible answer
   - an answer between 50 and 60

```
  [ ] [ ]
+ [ ] [ ]
---------
```

   How did you know what to choose?

2. Explain carrying to a second grader, using this problem:

```
   247
+  138
------
```

3. Jeff adds 62 and 73 on his calculator and gets 113. How do you know it's wrong?

4. Which is more and how do you know?

   324 or 432

   643 or 400 + 60 + 3

5. Sia's little sister wants to write two-hundred-forty-three like this: 20043. What would you tell her?

6. Find three two-digit numbers whose sum is 248.
Is there just one answer?
About how many answers are there?

7. Jo did a subtraction problem this way

$$\begin{array}{r} 425 \\ -\ 259 \\ \hline 234 \end{array}$$

Is Jo right or wrong?
What would you say to Jo?

8. Pick two numbers whose sum is 105 from this list:
36, 91, 54, 47, 30, 58
How did you do it?
Now you make up a problem like this one.

Cryptarithms

9. Replace each letter with one digit to make the example correct. The same letter gets the same digit each time it is used in one problem. Some problems might have more than one answer.

$$\begin{array}{r} TT \\ +\ VV \\ \hline WYW \end{array} \qquad\qquad \begin{array}{r} JK \\ +\ \ H \\ \hline LMM \end{array}$$

Tic Tac Toe (Problem solving, estimating, performing addition both mentally and with paper and pencil)

Need: 5 markers in each of two colors
Tic Tac Toe board below.

Take turns with a friend. Each of you chooses a color marker.
Pick the place where you want to put your marker.
Then pick two addends that you think will give you that sum.
You must put your marker on the sum of the addends you pick.
Three markers in a row of one color wins.

Addends: 17 23 45 32 28

49	68	40
55	62	45
73	77	51

Can you make up a tic tac toe board with different numbers and addends?
Try it to share with a friend.

Buggy Problems

What would you say about each child's work?

San	Jeremy	Jill	Jack	Jeff
57	57	57	57	57
+ 26	+ 26	+ 26	+ 26	+ 26
73	713	83	101	31

Appendix B

Coding Scheme

HB 9/23/93

Tentative Coding Scheme: Revised

know-m	What does it mean to know math)

instruction codes:

insgoals	(teachers' goals for mathematics learning and instruction)
insorg	(organization and management of instruction)
inswhat	(instructional tasks, activities, & materials; enacted curriculum)

assessment codes:

asgoals	(roles, goals and purposes for assessment)
ashow	(content/substance of assessment tasks; how teachers assess)
asscore	(scoring of assessment tasks)
track	(how to keep track of what students know)
grd	(how to assign grades in math)
as/rn	(what do you want to learn about assessment in this project)
tdil	(teacher dilemmas)
rdil	(researcher dilemmas)
student	(student knowledge, beliefs, attitudes, performances in mathematics)

advantages and limitations:

asadv	(advantages of performance assessments)
aslim	(limitations of performance assessments)
projadv	(advantages of the project)
projlim	(limitations of the project)

NOTE: Also indicate instances where teachers talk explicitly about change by using a *delta*. Double code these instances--once with the 'regular code" and once with the "delta code" E.g.,

delta-know-m & know-m for teacher's comments about changes in her ideas concerning what it means to know math

delta-aswhy & aswhy for teacher's reported changes in her ideas about the roles and purposes for assessment

Dimensions for Key Areas
Learning, Curriculum, and Instruction and Assessment in Mathematics

I. Beliefs and practice about how and what children learn

Direct instruction	Constructivist instruction
kids learn from being told	Kids figure things out themselves.
Memorizing is knowing.	Being able to use it is knowing.
Only some children can think mathematically.	All children can learn to think mathematically.
Children know their facts, procedures.	In addition, children can reason, solve problems, communicate.

II. Beliefs and practice about what school math is; what's important to learn, assess

Facts, computations, procedures, definitions, copying examples from text	Mathematical thinking, patterns, relationships, explanations
Math as the trivial, mechanical	Math as meaningful; making sense of math
Limited view of understanding	Extended view of understanding
Product	Process

III. Beliefs and practice about instruction and assessment

A. General

Uses textbook pages, worksheets; drill on facts, definitions, and computation	Uses worthwhile mathematical tasks that require thinking, reasoning, generalization, communication
T explains, shows how to do	T poses problems, asks questions, guides, orchestrates
Ss practice what they've been shown; memorize facts, definitions, procedures	Ss work on problems, discuss, report, question others

B. Problem solving

Story problems from text	Authentic, essential problems (everyday & mathematical)
Single answer	Open—multiple approaches, solutions
Well defined, very structured	Not well defined, unstructured
Contrived	Authentic
Only correct answer counts	Use of rubrics (criteria public); process values

C. Explanations

Not requested	Seen as important—both as a skill and as a window to mathematical thinking Ss asked to explain and justify solutions

D. Instruction/assessment materials

Textbook, worksheets	Tasks to demonstrate, solve, discuss
Limited use of manipulatives, calculators	Open use of manipulatives, calculators

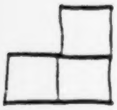
E. Additional Assessment Dimensions

Separate from instruction	could serve as good instruction; enhances instruction
Limited data—timed tests, chapter tests, comp'n tests	Multiple sources of data—problem solving, observations, alternative paper and pencil tasks
Gut feelings about students	Systematic records about students
Assessment of what Ss have been shown	Assessment requires extension and application
Learned nothing new about students	Learned significant new things about students
Doesn't assess activities, problem solving	Gets assessment information from non-p&p activities

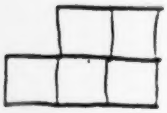
Appendix C

Examples of Teachers' Assessments

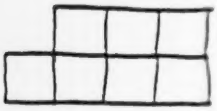
Building Buildings



1st building, 3 squares



2nd building, 5 squares



3rd building, 7 squares

How many squares are in the 8th building?

Building	Squares
1st	3
2nd	5
3rd	7

Explain

Name ____________ Date (Fall)

Subtraction Test

1. 48
-19

2. 658
-247

3. 428
-379

4. 700
-236

5. $4.06
-1.78

6. 7,163
-4,174

7. 1,600
- 782

Estimate

8. 78
-63

Design a problem that you would have to regroup.

9. ☐☐
-☐☐

Design a problem that would have an estimated answer of 30.

10. ☐☐
-☐☐

BEST COPY AVAILABLE

Name ______________________________ Date ________(Spring)________

Multiplication Assessment

1. Draw a picture that shows 3 X 7.

2. Show all the possible ways you could arrange 24 chairs in rows. Use "x" to symbolize a chair.

3. Use 3, 2, and 5. Make as many combinations that give products under 20. For example: 3 X 2= 6

4. How many legs do 7 cows have?

5. Write a multiplication story that is solved with 4 X 5. The story must end with a question.